

Inference for the Proportional Hazards Model with Misclassified Discrete-Valued Covariates

David M. Zucker

Department of Statistics, Hebrew University

Mount Scopus, Jerusalem, Israel

email: mszucker@mscc.huji.ac.il

and

Donna Spiegelman

Departments of Epidemiology and Biostatistics

Harvard School of Public Health

Boston MA, USA

November 20, 2003

SUMMARY

We consider the Cox proportional hazards model with discrete-valued covariates subject to misclassification. We present a simple estimator of the regression parameter vector for this model. The estimator is based on a weighted least squares analysis of weighted-averaged transformed Kaplan-Meier curves for the different possible configurations of the observed covariate vector. Optimal weighting of the transformed Kaplan-Meier curves is described. The method is designed for the case in which the misclassification rates are known or are estimated from an external validation study. A hybrid estimator for situations with an internal validation study is also described. When there is no misclassification, the regression coefficient vector is small in magnitude, and the censoring distribution does not depend on the covariates, our estimator has the same asymptotic covariance matrix as the Cox partial likelihood estimator. We present results of a finite-sample simulation study under Weibull survival in the setting of a single binary covariate with known misclassification rates. In this simulation study, our estimator performed as well as or, in a few cases, better than the full Weibull maximum likelihood estimator. We illustrate the method on data from a study of the relationship between trans-unsaturated dietary fat consumption and cardiovascular disease incidence.

KEY WORDS: Errors in variables, Kaplan-Meier curves, misclassification, survival regression

1. Introduction, Setup, and Notation

Beginning with Prentice (1982), a considerable literature has developed on survival analysis under the Cox (1972) proportional hazards model when the covariates are measured with error. The problem has been studied in various contexts. In this paper, we deal with a vector of discrete-valued covariates subject to misclassification. We present a simple estimator for the regression parameter vector taking the covariate misclassification into account. The misclassification is allowed to be arbitrarily complex. Basic statistical properties of the estimator are described. We consider the case where the classification rates are known or estimated from an external validation study. In addition, we present a simple hybrid estimator for situations involving an internal validation study. All of these procedures involve a combination of standard survival analysis methods with straightforward computations.

The setup is as follows. We assume that the survival time T^0 follows the proportional hazards model

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\psi(\mathbf{X};\boldsymbol{\beta}) \tag{1}$$

where $\lambda(t|\mathbf{X})$ represents the hazard function and \mathbf{X} is the p -vector of true covariate values. The function $\lambda_0(t)$ is a baseline hazard function of unspecified form. The function $\psi(\mathbf{x};\boldsymbol{\beta})$ represents the relative risk for an individual with covariate vector \mathbf{x} . This function involves a p -vector $\boldsymbol{\beta}$ of unknown parameters which are to be estimated. The classical Cox (1972) model assumes $\psi(\mathbf{x};\boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T\mathbf{x}}$. Following Thomas (1981) and Breslow and Day (1993, Sec. 5.1(c)), we allow a general relative risk function $\psi(\mathbf{x};\boldsymbol{\beta})$. The function $\psi(\mathbf{x};\boldsymbol{\beta})$ is assumed to be positive in a neighborhood of the true $\boldsymbol{\beta}$ for all possible configurations of \mathbf{x} and to be continuously differentiable with respect to the components of $\boldsymbol{\beta}$. We assume further that $\psi(\mathbf{x};\mathbf{0}) = 1$, which simply means that $\boldsymbol{\beta} = \mathbf{0}$ corresponds to no covariate effect. In many applications, it will be desirable to take $\psi(\mathbf{x};\boldsymbol{\beta})$ to be a function that is monotone in each component of \mathbf{x} for all $\boldsymbol{\beta}$. As usual, the data are subject to right censoring with censoring time U , and the observed survival data consist of the observed follow-up time $T = \min(T^0, U)$ and the event

indicator $\delta = I(T^0 \leq U)$.

We assume that the l -th component X_l of \mathbf{X} is discrete-valued with K_l numerical levels, so that there are $K_1 K_2 \cdots K_p$ total possible configurations of \mathbf{X} . We discard the configurations that have zero probability. We denote by K the number of remaining configurations, which we represent as $\mathbf{x}_1, \dots, \mathbf{x}_K$, in arbitrary order. It is assumed that $K \geq p + 1$. The covariates are subject to possible misclassification. In Secs. 2–4, we assume that we are unable to observe \mathbf{X} at all, but instead observe only an error-prone surrogate \mathbf{Z} . The range of possible values of \mathbf{Z} is assumed to be the same as that for \mathbf{X} . In particular, \mathbf{Z} is also discrete-valued. Both \mathbf{X} and \mathbf{Z} are time-independent. We define the classification rate matrix \mathbf{A} by $A_{rs} = \Pr(\mathbf{X} = \mathbf{x}_s | \mathbf{Z} = \mathbf{x}_r)$. This matrix is square. Initially we assume the A_{rs} are known. In Sec. 4, we discuss the case where the A_{rs} are estimated from an external validation study, that is, a separate, independent, study involving subjects with measurements on both \mathbf{X} and \mathbf{Z} but no survival data. In Sec. 5, we consider the main study / internal validation study design, which involves a main study where only \mathbf{Z} and the survival data (T, δ) are observed and an internal validation study where both \mathbf{X} and \mathbf{Z} along with (T, δ) are observed. We make the following assumptions throughout:

Assumption A: The random variables \mathbf{Z} and T^0 are conditionally independent given \mathbf{X} . This corresponds to Prentice (1982, Eqn. 2), and basically means that the measurement error mechanism is independent of the survival process.

Assumption B: The censoring time U is conditionally independent of T^0 given \mathbf{Z} . This is a modified version of the condition stated in Prentice (1982, Eqn. 5).

We work with n independent observations under the foregoing model, indexed by i . In developing our estimator, we partition the data set according to levels of the observed covariate vector \mathbf{Z} . We let n_k denote the number of observations with $\mathbf{Z} = \mathbf{x}_k$ and define $\pi_k = \Pr(\mathbf{Z} = \mathbf{x}_k)$. Our estimator is constructed from estimates of certain survival functions and cumulative hazard functions. We define here the relevant

notation for these quantities. We let $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ represent the cumulative hazard function corresponding to $\lambda_0(t)$. We denote the survival function, cumulative hazard function, and “at risk” function, respectively, for the set of observations with $\mathbf{Z} = \mathbf{x}_k$ by $S_k^*(t) = \Pr(T^0 > t | \mathbf{Z} = \mathbf{x}_k)$, $\Lambda_k^*(t) = -\log S_k^*(t)$, and $G_k^*(t) = \Pr(T \geq t | \mathbf{Z} = \mathbf{x}_k)$. The corresponding survival and cumulative hazard functions for subgroups of the population broken down by the value of the unobserved true covariate \mathbf{X} are denoted by $S_k(t) = \Pr(T^0 > t | \mathbf{X} = \mathbf{x}_k) = \exp(-\Lambda_0(t)\psi(\mathbf{x}_k; \boldsymbol{\beta}))$ and $\Lambda_k(t) = -\log S_k(t)$.

All asymptotic statements made in this paper pertain to the situation in which the sample size n tends to infinity while the number of covariates p and the number of covariate configurations K remains fixed.

2. The Estimator

As indicated by Prentice (1982), measurement error in the covariates creates difficulties in applying the classical Cox (1972) partial likelihood approach to estimation. Cox’s estimation procedure in its original form leads to biased estimates. Most procedures presented in the literature for the Cox model with measurement error involve adapting the Cox partial likelihood procedure to account for the measurement error. A variety of modified partial likelihood procedures have appeared in the literature, ranging from simple approximate procedures to more complex procedures. Here we attack the problem from a different angle, approaching the problem through the survival function.

In view of the relation $S_k(t) = \exp(-\Lambda_0(t)\psi(\mathbf{x}_k; \boldsymbol{\beta}))$, if we could estimate the S_k ’s then we could develop an estimator for $\boldsymbol{\beta}$. Now $S_k(t)$ cannot be estimated directly because the true \mathbf{X} values are unknown. On the other hand, $S_k^*(t)$ can be estimated because the \mathbf{Z} values are known. The first step, then, is to estimate the $S_k(t)$ from the $S_k^*(t)$.

Let $\mathbf{S}(t)$ and $\mathbf{S}^*(t)$ be K -vectors with components $S_k(t)$ and $S_k^*(t)$, respectively, and let \mathbf{B} be the matrix inverse of \mathbf{A} . The inverse will exist provided the misclassification

rates are not unduly extreme; see the discussion in Appendix A.1. By elementary probability manipulations, we have $\mathbf{S}^*(t) = \mathbf{A}\mathbf{S}(t)$, so that $\mathbf{S}(t) = \mathbf{B}\mathbf{S}^*(t)$. Now let $\hat{S}_k^*(t)$ be the Kaplan-Meier estimator of $S_k^*(t)$ based on the sample of subjects with $\mathbf{Z} = \mathbf{x}_k$, and let $\hat{\mathbf{S}}^*(t)$ be the vector comprising the $\hat{S}_k^*(t)$. Then we may estimate $\mathbf{S}(t)$ by $\hat{\mathbf{S}}(t) = \mathbf{B}\hat{\mathbf{S}}^*(t)$. Consequently, we may estimate $\Lambda_k(t)$ by $\hat{\Lambda}_k(t) = h(\hat{S}_k^*(t))$ with $h(s) = -\log s$.

Now $\Lambda_k(t) = \Lambda_0(t)\psi(\mathbf{x}_k; \boldsymbol{\beta})$, and thus based on $\hat{\Lambda}_k(t), k = 1, \dots, K$, for any fixed t we could construct an estimate of $\boldsymbol{\beta}$ using a nonlinear least squares procedure. It is, however, advantageous to pool information over the various values of t . We accomplish this pooling by constructing a weighted average of $\hat{\Lambda}_k(t)$ over t and applying nonlinear least squares to these weighted averages. We thus define

$$\hat{L}_k = \int_0^\infty \eta(t)\hat{\Lambda}_k(t)dt, \quad L_k = \int_0^\infty \eta(t)\Lambda_k(t)dt, \quad (2)$$

where $\eta(t)$ is a weight function integrating to one. The choice of the function $\eta(t)$ will be discussed in Sec. 3 below. We have

$$\hat{L}_k = L_k + \epsilon_k = e^{\beta_0}\psi(\mathbf{x}_k; \boldsymbol{\beta}) + \epsilon_k, \quad (3)$$

where $\beta_0 = \log \int \eta(t)\Lambda_0(t)dt$ and ϵ_k is a random error term. Let \mathbf{L} , $\hat{\mathbf{L}}$, and $\boldsymbol{\epsilon}$ denote vectors with components given by the quantities in (3). It is known from standard survival analysis theory (Breslow and Crowley, 1974; Gill, 1983) that the Kaplan-Meier estimator $\hat{S}_k^*(t)$ is a consistent estimator of $S_k^*(t)$ and that $\sqrt{n_k}\{\hat{S}_k^*(t) - S_k^*(t)\}$ converges to a Gaussian process. From these results, it follows that the asymptotic distribution of the vector $\sqrt{n}\boldsymbol{\epsilon}$ is mean-zero multivariate normal. In Appendix A.2, it is shown that the asymptotic covariance matrix $\boldsymbol{\Omega}$ of $\sqrt{n}\boldsymbol{\epsilon}$ is given by

$$\Omega_{rs} = \sum_{k=1}^K \pi_k^{-1} B_{rk} B_{sk} R_{rsk} \quad (4)$$

with

$$R_{rsk} = \int_0^\infty \left\{ \int_y^\infty \eta(t) \frac{S_k^*(t)}{S_r(t)} dt \right\} \left\{ \int_y^\infty \eta(u) \frac{S_k^*(u)}{S_s(u)} du \right\} G_k^*(y)^{-1} d\Lambda_k^*(y). \quad (5)$$

An estimate of Ω_{rs} may be obtained by replacing $S_k^*(t)$ by its Kaplan-Meier estimate, $\Lambda_k^*(t)$ by its Nelson-Aalen estimate, $G_k^*(t)$ by the proportion of subjects having follow-up time at least t among those with $\mathbf{Z} = \mathbf{x}_k$, and π_k by $\hat{\pi}_k = n_k/n$. At the end of the next section, a more explicit formula for R_{rsk} is provided under the weight function $\eta(t)$ that we ultimately propose.

Given the estimator $\hat{\Omega}$ of Ω , we may proceed with the procedure for estimating β . We propose to fit (3) by nonlinear weighted least squares. Let $\gamma = (\beta_0 \beta_1 \cdots \beta_p)^T$ and write $L_k(\gamma) = L_k = e^{\beta_0} \psi(\mathbf{x}_k; \beta)$. We estimate γ by the minimizer $\hat{\gamma}$ of the function $\Sigma(\gamma) = \{\hat{\mathbf{L}} - \mathbf{L}(\gamma)\}^T \hat{\Omega}^{-1} \{\hat{\mathbf{L}} - \mathbf{L}(\gamma)\}$, and then take our estimator $\hat{\beta}$ of β to be the vector formed by the last p elements of $\hat{\gamma}$. The estimation of γ may be carried out using standard nonlinear regression algorithms (Bates and Watts, 1988, Sec. 2.2; Seber and Wild, 1989, Sec. 2.1.4). In our software we use the Gauss-Newton method. Define $W_{kl}(\gamma) = \partial L_k(\gamma) / \partial \gamma_l$ for $k = 1, \dots, K$ and $l = 1, \dots, p + 1$. We have $\mathbf{W} = e^{\beta_0} \Psi$, where $\Psi_{k1} = \psi(\mathbf{x}_k; \beta)$ for $k = 1, \dots, K$ and $\Psi_{k,l+1} = \dot{\psi}_l(\mathbf{x}_k; \beta)$ for $k = 1, \dots, K$ and $l = 1, \dots, p$. Here $\dot{\psi}_l(\mathbf{x}; \beta)$ denotes the derivative of $\psi(\mathbf{x}; \beta)$ with respect to β_l . The estimating equation is given by

$$\mathbf{W}(\gamma)^T \hat{\Omega}^{-1} (\hat{\mathbf{L}} - \mathbf{L}(\gamma)) = \mathbf{0}, \quad (6)$$

and the Gauss-Newton iterations take the form

$$\gamma^{(r+1)} = \gamma^{(r)} + (\mathbf{W}^T \hat{\Omega}^{-1} \mathbf{W})^{-1} \mathbf{W}^T \hat{\Omega}^{-1} (\hat{\mathbf{L}} - \mathbf{L}(\gamma))|_{\gamma=\gamma^{(r)}}. \quad (7)$$

The asymptotic properties of the estimator γ may be developed using standard arguments for nonlinear regression (Jennrich, 1969; Seber and Wild, 1989, Chapter 12). In our setting, instead of the usual case with p and $\text{Var}(\epsilon_k)$ fixed and $K \rightarrow \infty$, we have a simpler case with p and K fixed and $\text{Var}(\epsilon_k) \rightarrow 0$. Appendix A.3 sketches the asymptotic argument. Under technical conditions stated in Appendix A.3, the estimator is consistent and asymptotically normal. The asymptotic covariance matrix of $\sqrt{n}(\hat{\gamma} - \gamma)$ is given by

$$\tilde{\mathbf{V}} = (\mathbf{W}^T \Omega^{-1} \mathbf{W})^{-1}, \quad (8)$$

and the asymptotic covariance matrix \mathbf{V} of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is the lower right $p \times p$ block of $\tilde{\mathbf{V}}$.

For the classical Cox model with $\psi(\mathbf{x}; \boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T \mathbf{x}}$, we can write $\mathbf{W}(\boldsymbol{\gamma})$ in the form $\mathbf{W}(\boldsymbol{\gamma}) = \mathbf{M}(\boldsymbol{\gamma})\boldsymbol{\mathcal{X}}$. Here $\mathbf{M}(\boldsymbol{\gamma})$ is a diagonal matrix with diagonal entries $L_k(\boldsymbol{\gamma})$, and $\boldsymbol{\mathcal{X}}$ is the matrix given by $\mathcal{X}_{k1} = 1$ for $k = 1, \dots, K$ and $\mathcal{X}_{k,l+1} = (\mathbf{x}_k)_l$ for $k = 1, \dots, K$ and $l = 1, \dots, p$. Also, when $\psi(\mathbf{x}; \boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T \mathbf{x}}$ a non-iterative alternate estimator is available. Let $Y_k = \log(L_k)$, and let $\mathbf{Y} = (Y_1 \dots Y_K)^T$. We then have $\mathbf{Y} = \boldsymbol{\mathcal{X}}\boldsymbol{\gamma} + \boldsymbol{\xi}$, where $\sqrt{n}\boldsymbol{\xi}$ has asymptotic covariance matrix \mathbf{C} given by $C_{rs} = \{L_r(\boldsymbol{\gamma})L_s(\boldsymbol{\gamma})\}^{-1}\hat{\Omega}_{rs}$. We may estimate \mathbf{C} by $\hat{C}_{rs} = \{\hat{L}_r\hat{L}_s\}^{-1}\Omega_{rs}$. We may then estimate $\boldsymbol{\gamma}$ by

$$\boldsymbol{\gamma}_{alt} = (\boldsymbol{\mathcal{X}}^T \hat{\mathbf{C}}^{-1} \boldsymbol{\mathcal{X}})^{-1} \boldsymbol{\mathcal{X}}^T \hat{\mathbf{C}}^{-1} \mathbf{Y}. \quad (9)$$

This estimator is asymptotically equivalent to the iterative estimator described above. When $K = p + 1$, both estimators are equal to $\boldsymbol{\mathcal{X}}^{-1}\mathbf{Y}$.

In the case of a single binary covariate, the estimate of the relative risk ψ is simply $\hat{\psi} = \hat{L}_2/\hat{L}_1$. We have

$$\text{Var}(\sqrt{n}(\hat{\psi} - \psi)) \rightarrow \frac{\sum_{k=1}^2 \pi_k^{-1} \int_0^\infty \{\int_s^\infty \eta(t) \varphi_k(t) S_k^*(t) dt\}^2 G_k^*(s)^{-1} d\Lambda_k^*(s)}{\{\int_0^\infty \eta(t) \Lambda_0(t) dt\}^2}, \quad (10)$$

where $\varphi_k(t) = \psi B_{1k} S_1(t)^{-1} - B_{2k} S_2(t)^{-1}$.

3. Optimal Weight Function

We now seek to identify the weight function $\eta(t)$ in (2) that is optimal in the sense of leading to an estimator of minimum variance. Because a general optimal solution is difficult to derive, we derive an optimal solution under the following simplifying assumptions:

W1. We have a “small $\boldsymbol{\beta}$ ” scenario with $\boldsymbol{\beta} = O(n^{-\frac{1}{2}})$, corresponding to what would be a local alternative in a hypothesis testing setting.

W2. The censoring distribution given $\mathbf{Z} = \mathbf{x}_k$ is the same for all k .

Later in this section we briefly discuss more general cases.

We stress that only the theoretical optimality of the weight function derived in this section depends on Assumptions (W1) and (W2) above. The consistency and asymptotic normality results presented in the previous section apply to the resulting estimator without any reliance on these assumptions. Thus, inferences based on our estimator are valid irrespective of whether Assumptions (W1) and (W2) hold. Moreover, it is reasonable to expect that our proposed estimator will have good efficiency for a range of β values beyond the “small β ” scenario, even though the estimator will not necessarily be exactly optimally efficient. This expectation is borne out by the simulation study presented in Sec. 6.

To proceed, we develop an expression for the asymptotic covariance matrix $\tilde{\mathbf{V}}$ of $\hat{\gamma}$ under Assumptions (W1) and (W2). As a preliminary step, we note that under (W1) and (W2) the following statements hold in the limit as $n \rightarrow \infty$.

1. The $S_k(t)$ and the $S_k^*(t)$ are all equal to $S_0(t) = e^{-\Lambda_0(t)}$.
2. The $G_k^*(t)$ are all equal to a common function $G(t)$.
3. A consequence of the foregoing two statements, the R_{rsk} 's in (5) are all equal to the common value

$$R = \int_0^\infty \left\{ \int_s^\infty \eta(t) dt \right\}^2 G(s)^{-1} d\Lambda_0(s).$$

4. The matrix \mathbf{W} defined at the end of Sec. 2 reduces to $e^{\beta_0} \Psi_0$, where Ψ_0 is Ψ evaluated at $\beta = \mathbf{0}$ (for $\psi(\mathbf{x}; \beta) = e^{\beta^T \mathbf{x}}$, we have $\Psi_0 = \mathcal{X}$).

Given the foregoing four statements, we arrive at the following conclusion: Define $\mathbf{\Pi}$ to be a diagonal matrix with diagonal elements π_k , and define

$$\rho = \frac{\int_0^\infty \left\{ \int_s^\infty \eta(t) dt \right\}^2 G(s)^{-1} d\Lambda_0(s)}{\left\{ \int_0^\infty \eta(s) \Lambda_0(t) ds \right\}^2}. \quad (11)$$

Then, under (W1) and (W2), as $n \rightarrow \infty$ we have $\mathbf{\Omega} = \rho \mathbf{B} \mathbf{\Pi}^{-1} \mathbf{B}^T$ and thus $\tilde{\mathbf{V}} = \rho e^{-2\beta_0} (\mathbf{\Psi}_0^T \mathbf{A}^T \mathbf{\Pi} \mathbf{A} \mathbf{\Psi}_0)^{-1}$.

To find the optimal $\eta(t)$, we need to minimize ρ . In Appendix A.4, it is shown that choice of $\eta(t)$ that minimizes ρ is given by $\eta(t)dt = -dG(t)$. In practice, we replace $-dG(t)$ by the measure that puts a point mass of n^{-1} at each observed follow-up time T_i irrespective of whether individual i had an observed event or was censored. In other words, we define

$$\hat{L}_k = \frac{1}{n} \sum_{i=1}^n \hat{\Lambda}_k(T_i). \quad (12)$$

For the case of a single binary covariate ($K=2$) with possibly different censoring distributions for the two levels of Z , a similar argument may be used to identify the optimal weight function for the “small β ” scenario. The result is $\eta(t)dt = -d\tilde{G}(t)$, where $\tilde{G}(t) = \{\pi_0^{-1}G_1^*(t)^{-1} + \pi_1^{-1}G_2^*(t)^{-1}\}^{-1}$. The functions $G_k^*(t)$ may be estimated as indicated in the preceding section. For general β , it does not appear possible to solve analytically for the optimal weight function. However, an approximate numerical solution may be obtained by discretizing the integrals in the variance expression (10) and minimizing the resulting approximate variance expression by quadratic programming.

For $K > 2$ and different censoring distributions for the various levels of \mathbf{Z} , the situation is considerably more complex. We have a problem involving optimization of a covariance matrix. Problems of this type have no unique solution; experimental design researchers have proposed various optimality criteria (Steinberg and Hunter, 1984, Sec. 3). In our case the process of finding the weighting that optimizes these criteria appears rather involved. We therefore leave the problem as a possible topic for further work. As an approximate solution, we propose simply using the optimal weight function derived for the “small β ” scenario under the assumption of a common censoring distribution.

The expression (4) for the covariance matrix $\mathbf{\Omega}$ involves the weight function $\eta(t)$ and the functions $G_k^*(t)$. Our proposed choice for $\eta(t)$ is based on the assumption of

a common censoring distribution. However, the functions $G_k^*(t)$ should be estimated separately for each k in the manner indicated in the previous section. Ultimately, we estimate the R_{rsk} in (5) by

$$\hat{R}_{rsk} = \sum_{i: \mathbf{Z}_i = \mathbf{x}_k} Q_{rk}(T_i) Q_{sk}(T_i) \hat{G}_k^*(T_i)^{-1} (\hat{\Lambda}_k^*(T_i) - \hat{\Lambda}_k^*(T_i-)), \quad (13)$$

where

$$Q_{rk}(t) = \frac{1}{n} \sum_{j: T_j \geq t} \frac{\hat{S}_k^*(T_j)}{\hat{S}_r(T_j)}. \quad (14)$$

4. Main Study / External Validation Study Design

We now develop the theory for the case where the A_{rs} are estimated from an external validation study with sample size m . This theory is applied in the example presented in Sec. 7. In general, we may express \mathbf{A} as $\mathbf{A}(\boldsymbol{\omega})$ for some q -vector of parameters $\boldsymbol{\omega}$. The nature of the function $\mathbf{A}(\boldsymbol{\omega})$ is dictated by the measurement error model employed. For example, consider the case of two 0–1 binary covariates, so that $x_1 = (00)^T$, $x_2 = (10)^T$, $x_3 = (01)^T$, and $x_4 = (11)^T$. As an illustration, suppose that the covariates X_1 and X_2 are statistically independent and are both subject to misclassification of unspecified structure. Assume further that the error processes for the two covariates are independent of each other. Then

$$\mathbf{A}(\boldsymbol{\omega}) = \begin{bmatrix} \omega_1 \omega_3 & (1 - \omega_1) \omega_3 & \omega_1 (1 - \omega_3) & (1 - \omega_1) (1 - \omega_3) \\ (1 - \omega_2) \omega_3 & \omega_2 \omega_3 & (1 - \omega_2) (1 - \omega_3) & \omega_2 (1 - \omega_3) \\ \omega_1 (1 - \omega_4) & (1 - \omega_1) (1 - \omega_4) & \omega_1 \omega_4 & (1 - \omega_1) \omega_4 \\ (1 - \omega_2) (1 - \omega_4) & \omega_2 (1 - \omega_4) & (1 - \omega_2) \omega_4 & \omega_2 \omega_4 \end{bmatrix}. \quad (15)$$

This setup includes the case where X_1 is measured with error while X_2 is measured perfectly; we just set $\omega_3 = \omega_4 = 1$. If dependence between the two covariates or dependence in the error processes associated with the two covariates is present, the matrix $\mathbf{A}(\boldsymbol{\omega})$ will have a more complex structure involving additional parameters ω_r . Often, it will be appropriate to use a “saturated model” with a distinct parameter for each off-diagonal element of \mathbf{A} .

The external validation study is presumed to provide an estimator $\hat{\boldsymbol{\omega}}$ having an approximate normal distribution, with mean $\boldsymbol{\omega}$ and covariance matrix $m^{-1}\boldsymbol{\Gamma}$, along

with an estimator of the matrix $\mathbf{\Gamma}$. For example, for the case of a single 0–1 binary covariate, the estimates of $\omega_k = \Pr(X = k - 1 | Z = k - 1)$, $k = 1, 2$, are given by the obvious sample proportions. In this case, $\mathbf{\Gamma}$ is a 2×2 diagonal matrix with $\Gamma_{kk} = \omega_k(1 - \omega_k)/\vartheta_k$, where ϑ_k is the probability that $Z = k - 1$ in the external validation study. For the asymptotics, we assume that m and n are of the same order of magnitude, i.e., $m/n \rightarrow \zeta$ for some constant ζ as $n \rightarrow \infty$. Otherwise the error in $\mathbf{A}(\boldsymbol{\omega})$ will either be dominated by or will dominate the error in $\hat{\boldsymbol{\beta}}$ due to estimation of the S_k^* . Typically, ζ will be between 0 and 1.

We have to extend the Taylor expansion used in Appendix A.2 to account for the error in the A_{rs} 's. Let $\dot{\mathbf{A}}_r(\boldsymbol{\omega})$ denote the partial derivative of $\mathbf{A}(\boldsymbol{\omega})$ with respect to ω_r . Then, by the rule for differentiating an inverse matrix, the partial derivative of \mathbf{B} with respect to ω_r is $-\mathbf{B}\dot{\mathbf{A}}_r(\boldsymbol{\omega})\mathbf{B}$. Next, let $\hat{\mathbf{L}}(t)$, $\mathbf{L}(t)$, $\mathbf{D}(t)$, $\mathbf{H}(t)$, and $\tilde{\mathbf{H}}(t)$ be as defined in Appendix A.2. Then, extending Eqn. (17) of Appendix A.2, we may write

$$\hat{\mathbf{L}}(t) \doteq \mathbf{L}(t) + \mathbf{H}(t)(\hat{\mathbf{S}}^*(t) - \mathbf{S}^*(t)) + \boldsymbol{\phi}(t)(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}),$$

where the r -th *column* of the matrix $\boldsymbol{\phi}(t)$ is given by

$$\phi_r(t) = -\mathbf{D}(t)\mathbf{B}\dot{\mathbf{A}}_r(\boldsymbol{\omega})\mathbf{B}\mathbf{S}^*(t) = -\mathbf{D}(t)\mathbf{B}\dot{\mathbf{A}}_r(\boldsymbol{\omega})\mathbf{S}(t).$$

Accordingly,

$$\hat{\mathbf{L}} \doteq \mathbf{L} + \int_0^\infty \tilde{\mathbf{H}}(t)(\hat{\mathbf{S}}^*(t) - \mathbf{S}^*(t)) dt + \boldsymbol{\Phi}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}),$$

with $\boldsymbol{\Phi} = \int_0^\infty \eta(u)\boldsymbol{\phi}(u)du$. Thus, the term that must be added to the matrix $\boldsymbol{\Omega}$ to account for the error in estimating \mathbf{A} is $\zeta^{-1}\boldsymbol{\Phi}\mathbf{\Gamma}\boldsymbol{\Phi}^T$. The $\phi_r(t)$ may be estimated by substituting in the estimates for $\boldsymbol{\omega}$ and $\mathbf{S}(t)$. With the weighting $\eta(t)dt = -dG(t)$ proposed in the preceding section, we estimate $\boldsymbol{\Phi}$ by $\hat{\boldsymbol{\Phi}} = n^{-1} \sum_{i=1}^n \hat{\boldsymbol{\phi}}(T_i)$.

In regarding to the optimal weight function $\eta(t)$ in the “small $\boldsymbol{\beta}$ ” scenario, estimation of the A_{rs} turns out to have no effect. Under the “small $\boldsymbol{\beta}$ ” scenario, we have $\mathbf{S}(t) = S_0(t)\mathbf{e}$ and $\mathbf{D}(t) = d(t)\mathbf{I}$, where \mathbf{e} is the vector of ones and \mathbf{I} is the identity matrix. Thus, $\phi_r(t) = -d(t)S_0(t)\mathbf{B}\dot{\mathbf{A}}_r(\boldsymbol{\omega})\mathbf{e}$. Now $\mathbf{A}(\boldsymbol{\omega})\mathbf{e} = \mathbf{e}$ for all $\boldsymbol{\omega}$, because

$\sum_{l=1}^K \Pr(\mathbf{X} = \mathbf{x}_l | \mathbf{Z} = \mathbf{x}_k) = 1$ for all k . Therefore, $\dot{\mathbf{A}}_r(\boldsymbol{\omega})\mathbf{e} = \mathbf{0}$ identically and thus $\boldsymbol{\phi}_r(t) = \mathbf{0}$.

A FORTRAN program for implementing our method under the Cox (1972) model with $\psi(\mathbf{x}; \boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T \mathbf{x}}$, including allowance for uncertainty in the estimated classification rates, is available at the website <http://pluto.mscc.huji.ac.il/~mszucker>.

5. Main Study / Internal Validation Study Design

Here we develop theory for the case where the original study of n participants consists of two independent subsets: an internal validation study, in which both \mathbf{X} and \mathbf{Z} are measured, and a typically larger main study, in which only \mathbf{Z} is measured. This type of design is of substantial interest when covariate error is anticipated, and has been discussed by a number of authors (e.g., Zhou and Pepe, 1995; Carroll, Ruppert, and Stefanski, 1995; Spiegelman and Gray, 1991). For instance, Zhou and Pepe discussed a heart failure survival trial with this type of design. Ejection fraction, an important covariate, was measured in this trial using an error-prone non-standardized method in all patients and using a more accurate standardized method in a subset of patients. Zhou and Pepe presented an estimator of $\boldsymbol{\beta}$ based on a modified version of the Cox partial likelihood.

We propose a simple hybrid estimator based on a stratified analysis, with the main study and internal validation study regarded as strata. Spiegelman, Carroll, and Kipnis (2001) used a similar approach to combine an estimate from the internal validation study with a regression calibration estimate from the main study. The most common setup for a main study / internal validation study design is one in which sampling for inclusion in the internal validation study is completely random. It is also possible to consider sampling schemes based on \mathbf{Z} . If the probability of being sampled is a function of \mathbf{Z} only, then the classification probabilities $\Pr(\mathbf{X} = \mathbf{x}_s | \mathbf{Z} = \mathbf{x}_r)$ will be the same for the sampled and unsampled subjects. In this case the estimate of \mathbf{A} obtained from the internal validation sample will be appropriate for the remaining subjects in the main study.

For the internal validation study, we compute the standard Cox partial likelihood estimate of β . For the main study (omitting participants in the internal validation sample), we compute an estimate of β using the foregoing Kaplan-Meier based procedure. Here we use an estimate of \mathbf{A} based on the estimated joint distribution of \mathbf{X} and \mathbf{Z} obtained from the internal validation study. Call the resulting coefficient vector estimates $\hat{\beta}_M$ and $\hat{\beta}_{IV}$, respectively. Now the conditional expectation of $\hat{\beta}_{IV} - \beta$ given the \mathbf{X}_i and \mathbf{Z}_i values in the internal validation study is asymptotically zero. Thus, $\hat{\beta}_{IV} - \beta$ is asymptotically uncorrelated with the \hat{A}_{rs} . Because the random variables in question are asymptotically jointly normally distributed, the asymptotic zero correlation implies asymptotic independence.

It follows that the estimators $\hat{\beta}_{IV}$ and $\hat{\beta}_M$ from the two sub-studies are asymptotically independent. The covariance matrix \mathbf{V}_{IV} of $\hat{\beta}_{IV}$ is given by standard Cox model theory. For the usual model with $\psi(\mathbf{x}; \beta) = e^{\beta^T \mathbf{x}}$, an estimate of \mathbf{V}_{IV} may be obtained with Cox model software available in common statistical packages. For general relative risk functions, more specialized software must be used. The covariance matrix \mathbf{V}_M of $\hat{\beta}_M$ is as developed above. Our proposed overall estimator of β is an inverse-variance weighted average of the estimates from the two sub-studies: $\hat{\beta}_{Overall} = (\hat{\mathbf{V}}_{IV}^{-1} + \hat{\mathbf{V}}_M^{-1})^{-1}(\hat{\mathbf{V}}_{IV}^{-1}\hat{\beta}_{IV} + \hat{\mathbf{V}}_M^{-1}\hat{\beta}_M)$. This estimator will be asymptotically normal with mean β and covariance $(\mathbf{V}_{IV}^{-1} + \mathbf{V}_M^{-1})^{-1}$.

6. Simulation Study

To investigate the performance of our method, we performed a simulation study in the setting of a single 0–1 binary covariate with measurement error rates known. The simulation scenarios were constructed to represent situations typical of cancer epidemiology studies. The covariate represents presence or absence of some risk factor under investigation.

We ran two sets of simulations, both relating to studies of 5-year duration. The first set had a total sample size of 2,000 and a 5-year cumulative incidence rate of 25%

among those unexposed to the risk factor. The second set had a total sample size of 10,000 and a 5-year cumulative incidence rate of 5% among the unexposed. The baseline survival distribution was taken to be Weibull, with baseline hazard function $\lambda_0(t) = \theta\nu(\nu t)^{\theta-1}$. We took the power parameter θ equal to 5, which is typical of many types of cancer (Armitage and Doll, 1961; Breslow and Day, 1993, Sec. 6.3). The scale parameter ν was chosen so as to yield the specified cumulative incidence rate for the unexposed population. Censoring was taken to be exponential with a rate of 1% per year. For brevity of presentation, we take the false positive rate $\Pr(Z = 1|X = 0)$ and the false negative rate $\Pr(Z = 0|X = 1)$ to be equal to a common classification error rate. A range of values was used for the prevalence of the risk factor (5%, 25%, 40%), the classification error rate (1%, 5%, 10%, 20%), and the true relative risk (1.5, 2.0).

For comparison, we also present results for (a) the naive Cox (1972) partial likelihood estimator ignoring the measurement error and (b) the parametric maximum likelihood estimator (MLE) based on the full Weibull log likelihood under the relevant measurement error model. The Weibull log likelihood is given by

$$\begin{aligned} \ell = & \sum_{i=1}^n \delta_i \left\{ \log(\theta\nu(\nu T_i)^{\theta-1}) + \log \left(\frac{A(X_i + 1, 1)e^{-(\nu T_i)^\theta} + A(X_i + 1, 2)e^{e^\beta e^{-(\nu T_i)^\theta}}}{A(X_i + 1, 1)e^{-(\nu T_i)^\theta} + A(X_i + 1, 2)e^{-e^\beta(\nu T_i)^\theta}} \right) \right\} \\ & + \sum_{i=1}^n \log(A(X_i + 1, 1)e^{-(\nu T_i)^\theta} + A(X_i + 1, 2)e^{-e^\beta(\nu T_i)^\theta}), \end{aligned}$$

where we have written $A(r, s)$ for A_{rs} . The parametric MLE is of course an asymptotically optimally efficient estimator and therefore serves as a natural benchmark. In principle, an alternative estimator such as ours can match or approximate the asymptotic efficiency of the the parametric MLE. An alternate estimator also can possibly outperform the parametric MLE in a given finite sample situation. In all simulations, there were 5,000 replications.

Tables 1 and 2 summarize the simulation results for the $n = 2,000$ simulation set and the $n = 10,000$ simulation set, respectively. The following results are given: (a) the percent bias of the estimated log relative risk relative to the true value for the naive Cox estimator, our estimator, and the parametric Weibull estimator, (b) the

empirical variance of both our estimator and the Weibull estimator, and (c) the ratio of the mean square error of the Weibull estimator to that of our estimator, and (d) the empirically-estimated coverage probability of the normal-theory 95% confidence interval for the log relative risk based on our estimator.

We discuss first the results under an exposure prevalence of 25% or 40%. Here, in both simulation sets, the Cox estimator was substantially biased except under a misclassification rate of 1%. By contrast, our estimator exhibited excellent performance, including minimal bias in the estimation of the log relative risk and accurate confidence interval coverage. Our estimator and the Weibull estimator performed comparably in terms of bias and variance.

We turn now to the results under an exposure prevalence of 5%. Here, the performance of all three estimators was degraded, which is not surprising. In these cases, the expected number of events in the exposed group is only on the order of 25–50. Additionally, with an exposure prevalence of 5% and a misclassification rate of 5% or more, the predictive value of an observed positive exposure is low. These factors can be expected to lead to degraded performance. The naive Cox estimator was drastically biased. Our estimator and the Weibull estimator were dramatically less biased, though they did exhibit some degree of bias. The bias of our estimator and the Weibull estimator were generally comparable, with our estimator being noticeably superior in a few cases.

Overall, in terms of mean square error, the performance of our estimator in both simulation sets typically was essentially identical to that of Weibull estimator. In a few cases, our estimator was better. This finding exemplifies the fact that in a finite sample setting, the parametric MLE sometimes can be outperformed by an alternate estimator.

7. Example

As an illustrative example, we apply our method to data from the Nurses Health Study a prominent large-scale epidemiological study. Paralleling one of the analyses

presented by Hu et al. (1997), we examine the relationship between average daily trans-unsaturated fatty acid (TFA) consumption (g/day) and cardiovascular disease incidence. We use the classical Cox relative risk function $\psi(\boldsymbol{\beta}; \mathbf{x}) = e^{\boldsymbol{\beta}^T \mathbf{x}}$. The data consist of observations on 80,052 female nurses who underwent dietary assessment using a food frequency questionnaire (FFQ) in 1980 and were followed up to June 1, 1994 for cardiovascular events. Initial analyses examined the level of risk associated with each TFA quintile. It was found that there was an appreciable increased risk in the 5th quintile relative to the 1st quintile, but minimal increased risk in the 2nd-4th quintiles. Therefore, in the present analysis, we work with a binary “high TFA” risk factor defined as 1 for subjects in the 5th TFA quintile and 0 for the others. We also adjust for age in 1980 via strata defined as < 45 , $45-49$, $50-54$, $55+$. In the analyses presented by Hu et al., the results with adjustment for other covariates aside from age were similar to those with adjustment for age alone. Therefore here we consider only age adjustment. Thus, our model will have four regression coefficients, one for the binary risk factor (X_1) and three dummy variables for the age strata (X_2, X_3, X_4). There are 8 possible covariate configurations.

It is well known that the FFQ measures dietary intake with some degree of error and more reliable information can be obtained from a diet record (DR) (Willett, 1998, Chapter 6). We thus take X_1 to be the TFA risk factor indicator based on the DR and Z_1 to be the TFA risk factor indicator based on the FFQ. We estimate the elements of the classification matrix \mathbf{A} using data from the Nurse’s Health Study validation study (Willett et al., 1985). The validity of TFA intake calculated from the FFQ’s was not assessed directly in comparison with diet records because food composition databases for the latter are not available for period during which these studies were conducted. For this example, we therefore used the estimates of validity for saturated fat calculated from the same FFQ. The validity of the FFQ is likely to be somewhat greater for saturated fat than for TFA because saturated fat content depends less on processing methods that are subject to change over time. However, calculated TFA does correlate reasonably with adipose TFA (London et al., 1991; Hunter et al.,

1992), indicating that using the values for saturated fat is not unrealistic. The relevant estimated classification probabilities thus obtained were $\Pr(X_1 = 0|Z_1 = 0) = 0.84$ and $\Pr(X_1 = 1|Z_1 = 1) = 0.34$, with corresponding estimated standard errors of 0.031 and 0.080. The age strata dummy variables X_2, X_3, X_4 are assumed to be measured without error.

Table 4 presents the results of the following analyses: (1) a classical Cox regression analysis ignoring measurement error, (2) our method with \mathbf{A} set to be the identity matrix, corresponding to an assumption that there is no measurement error, (3) our method with \mathbf{A} assumed known and set according to the foregoing estimated classification probabilities, ignoring the estimation error in these probabilities, and (4) our method with \mathbf{A} estimated as above with the estimation error in the probabilities taken into account (main study / external validation study design). This last approach makes use of the theory developed in Sec. 4. The matrix \mathbf{A} has a structure similar to that indicated in (15), though here there are four covariates rather than just two. The ω 's corresponding to the age group indicator variables are all equal to one.

As expected, the results obtained with our method with the identity classification matrix were very similar to those obtained with the Cox analysis. The estimated relative risk associated with high TFA was 1.3, with 95% confidence interval ranging from 1.1 to 1.5. The relative risk estimate changed noticeably after correction for measurement error, which is unsurprising given the estimated level of misclassification. As regards the age effects, it is known in theory that measurement error in one covariate may affect the estimation of other covariate effects (Carroll, Ruppert, and Stefanski, 1995, Sec. 2.2.3). This phenomenon can occur when there are correlations among the underlying true covariates or correlations among the errors. In this example, the estimated coefficients for X_2 and X_3 were not substantially affected by the correction. The estimated coefficient for X_4 was affected to a noticeable degree under the non-iterative estimation procedure but minimally under the iterative estimation procedure. It appears that the iterative estimation procedure may be more stable.

The estimated relative risk for high TFA varied between the non-iterative and iterative method and between the analyses assuming known misclassification rates and those accounting for the estimation error in the misclassification rates. The estimated relative risk for high TFA ranged from 1.6 to 2.9 depending on the specific analysis, as compared with the value 1.3 obtained without measurement error adjustment. The standard errors of the adjusted log relative risk estimates were greater than those for the unadjusted estimates. This increase was greater when the estimation error in the misclassification rates was taken into account. With the estimation error in the misclassification rates taken into account, the 95% confidence interval for the relative risk associated with high TFA was 0.7 to 4.5 according to the non-iterative method and 0.7 to 4.0 according to the iterative method. Overall, it appears that the unadjusted relative risk estimate may be an underestimate, but the extent to which this so is uncertain.

9. Summary and Discussion

We have presented a simple estimator of the regression coefficient vector in the Cox proportional hazards survival model (1) with discrete-valued covariates subject to misclassification. We allow a general relative risk function $\psi(\mathbf{x}; \boldsymbol{\beta})$. We review here the main elements of our approach. Recall that \mathbf{X} denotes the true covariate vector, while \mathbf{Z} denotes the surrogate. We use the Kaplan-Meier estimates of $S_k^*(t) = \Pr(T^0 > t | \mathbf{Z} = \mathbf{x}_k)$ and an estimate of the classification rate matrix \mathbf{A} to obtain an estimate of $S_k(t) = \Pr(T^0 > t | \mathbf{X} = \mathbf{x}_k)$. We define $\hat{\Lambda}_k(t) = -\log \hat{S}_k(t)$. We then form a weighted average \hat{L}_k of $\hat{\Lambda}_k(t)$ as indicated in (2). Guided by theoretical optimality considerations, we take $\hat{L}_k = n^{-1} \sum_{i=1}^n \hat{\Lambda}_k(T_i)$. When \mathbf{A} is known, the asymptotic covariance matrix $\boldsymbol{\Omega}$ of the normalized \hat{L}_k is given by (4). This matrix is estimated using the substitutions indicated just below (4) and in (13) and (14). When \mathbf{A} is estimated from an external validation study, the estimate of $\boldsymbol{\Omega}$ must be adjusted as indicated at the end of Sec. 4. Given $\hat{\boldsymbol{\Omega}}$, the relevant parameter vector $\boldsymbol{\gamma} = (\beta_0 \beta_1 \cdots \beta_p)^T$ is estimated using the nonlinear least squares procedure defined by (6) and (7). Here

β_1, \dots, β_p are the regression coefficients of interest and β_0 is a subsidiary intercept parameter. The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by the lower $p \times p$ block of the matrix (8). For the classical Cox model with $\psi(\mathbf{x}; \boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T \mathbf{x}}$, the non-iterative alternate estimator (9) is also available. In Sec. 5, we describe a hybrid procedure for the main study / internal validation study design.

Thus, our approach can handle either internal or external validation study designs. It can handle arbitrarily complex measurement error structure, thus going beyond the independent additive error model often assumed. The method yields a consistent estimator. Because our approach is based on separate Kaplan-Meier curves for each of the K possible covariate configurations, it is probably necessary to have a reasonable amount of data at a reasonable fraction of the configurations in order for the approach to perform well. This proviso puts a certain limitation on the applicability of the approach. This issue can be dealt with to some degree by smoothing. Similarly, though we have developed our methods for discrete-valued covariates, in principle our approach can be extended to handle continuous-valued covariates by nonparametric smoothing. For instance, the Zhou and Pepe (1995) method for discrete-valued covariates was extended to continuous-valued covariates by Zhou and Wang (2000) using kernel smoothing. A similar approach could be used to extend our method to handle continuous-valued covariates. We have been engaged in further research on alternate methods aimed at overcoming the limitations of the method presented here, and we hope to report on these methods in due course.

For the case of a single perfectly measured binary covariate, Andersen (1983) and Kalbfleisch and Prentice (1981) presented simple relative risk estimators that are similar in spirit to ours, but different in form. Andersen showed the optimal version of his estimator to be asymptotically equivalent to the Cox (1972) partial likelihood estimator in the “small $\boldsymbol{\beta}$ ” scenario. Kalbfleisch and Prentice presented numerical efficiency comparisons between their estimators and the Cox estimator, but no theoretical efficiency results.

We have examined the asymptotic covariance matrix \mathbf{V} for our estimator in the standard Cox model with $\psi(\mathbf{x}; \boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T \mathbf{x}}$. Under the “small $\boldsymbol{\beta}$ ” scenario without covariate error and a common censoring distribution for the different levels of \mathbf{X} , the matrix \mathbf{V} is equal to ρ^* times the lower right $p \times p$ block of the matrix $(\boldsymbol{\mathcal{X}}^T \boldsymbol{\Pi} \boldsymbol{\mathcal{X}})^{-1}$, where ρ^* is the expression on the right hand side of Appendix Eqn. (19). We obtain $\mathbf{V} = \rho^* \boldsymbol{\Upsilon}^{-1}$, where

$$\boldsymbol{\Upsilon} = \sum_{k=1}^K \pi_k \mathbf{x}_k \mathbf{x}_k^T - \left\{ \sum_{k=1}^K \pi_k \mathbf{x}_k \right\} \left\{ \sum_{k=1}^K \pi_k \mathbf{x}_k \right\}^T.$$

It is easily seen that, under the assumptions just described, the standard expression for the covariance matrix of the Cox estimator reduces to the foregoing \mathbf{V} . A similar result holds when $K = 2$ without the common censoring assumption. For this case, the variance of the Cox estimator is given by Andersen’s (1983) Eqn. (9) with his θ_0 set equal to 1, and the variance of our estimator is equal to the same expression.

Thus, we obtain the following theoretical result: when there is no misclassification, $\boldsymbol{\beta}$ is small, and we have either $K = 2$ (i.e., a single binary covariate) or $K > 2$ with approximately equal censoring distributions at the different covariate configurations, our estimator is approximately as efficient as the Cox estimator. This result is interesting in its own right. Moreover, this result suggests that our estimator can be expected to be reasonably efficient in the presence of misclassification in comparison with other possible estimators. Indeed, in our finite-sample simulations for the case of a single binary covariate, our estimator was found to be essentially equivalent in efficiency to the fully parametric estimator in most of the cases. In a few cases, our estimator was actually somewhat better.

ACKNOWLEDGEMENT

This work was supported in part by U.S. National Cancer Institute Grant NCI CA50597.

APPENDIX

A.1. Remarks on the Invertibility of the Matrix \mathbf{A}

We discuss invertibility of \mathbf{A} when the error processes associated with the different covariates are independent of each other. Specifically, the assumption is that

$$\Pr(Z_l = x_{s_l}^{(l)}, l = 1, \dots, p | X_l = x_{r_l}^{(l)}, l = 1, \dots, p) = \prod_{l=1}^p \tilde{A}_{r_l s_l} \quad (16)$$

where $x_k^{(l)}, l = 1, \dots, p, k = 1, \dots, K_l$ represent the possible values of X_l and $\tilde{A}_{r_l s_l}^{(l)} = \Pr(Z_l = x_{s_l}^{(l)} | X_l = x_{r_l}^{(l)})$. Recall $\mathbf{\Pi} = \text{diag}(\pi_1 \dots \pi_K)$. Define $\tilde{\mathbf{\Pi}} = \text{diag}(\tilde{\pi}_1 \dots \tilde{\pi}_K)$ with $\tilde{\pi}_s = \Pr(\mathbf{X} = \mathbf{x}_s)$. Then, under (16), $\mathbf{A} = \tilde{\mathbf{\Pi}}(\tilde{\mathbf{A}}^{(1)} \otimes \tilde{\mathbf{A}}^{(2)} \otimes \dots \otimes \tilde{\mathbf{A}}^{(p)})\mathbf{\Pi}^{-1}$, with \otimes denoting the Kronecker product. Thus \mathbf{A} will be invertible if all of the $\tilde{\mathbf{A}}^{(l)}$ are invertible (Rao, 1973, p. 29). A sufficient condition for invertibility of a matrix \mathbf{F} is that $F_{rr} > \sum_{s \neq r} F_{rs}$ (Horn and Johnson, 1985, p. 349). The matrix $\tilde{\mathbf{A}}^{(l)}$ will satisfy this condition if $\Pr(Z_l = x_k^{(l)} | X_l = x_k^{(l)}) > \sum_{r \neq k} \Pr(Z_l = x_r^{(l)} | X_l = x_k^{(l)})$ for each k and l . This latter condition is one that may be expected to hold for any reasonable set of surrogate covariates.

A.2. The Matrix $\mathbf{\Omega}$

We develop here an expression for $\mathbf{\Omega}$. Let \doteq signify equality up to negligible terms. By Taylor expansion we have

$$\hat{\Lambda}_k(t) \doteq \Lambda_k(t) + d_k(t)(\hat{S}_k(t) - S_k(t)),$$

where $d_k(t) = h'(S_k(t)) = -S_k(t)^{-1}$. Now let $\mathbf{L}(t)$ denote the vector comprising the $\Lambda_k(t)$, and let $\hat{\mathbf{L}}(t)$ the vector comprising the $\hat{\Lambda}_k(t)$. In addition, let $\mathbf{D}(t)$ denote a diagonal matrix with diagonal elements $d_k(t)$, and define $\mathbf{H}(t) = \mathbf{D}(t)\mathbf{B}$. We then have

$$\hat{\mathbf{L}}(t) \doteq \mathbf{L}(t) + \mathbf{D}(t)\mathbf{B}(\hat{\mathbf{S}}^*(t) - \mathbf{S}^*(t)) = \mathbf{L}(t) + \mathbf{H}(t)(\hat{\mathbf{S}}^*(t) - \mathbf{S}^*(t)). \quad (17)$$

Therefore

$$\hat{\mathbf{L}} \doteq \mathbf{L} + \int_0^\infty \tilde{\mathbf{H}}(t)(\hat{\mathbf{S}}^*(t) - \mathbf{S}^*(t)) dt,$$

where $\tilde{\mathbf{H}}(t) = \eta(t)\mathbf{H}(t)$. Now define $\Delta_k(t) = \sqrt{n_k}\{\hat{S}_k^*(t) - S_k^*(t)\}$. By the asymptotic form of Greenwood's formula (Breslow and Crowley, 1974, Eqn. 6.3; Gill, 1983, p. 50),

$$\text{Cov}(\Delta_k(t), \Delta_k(u)) \doteq S_k^*(t)S_k^*(u) \int_0^{\min\{t,u\}} G_k^*(s)^{-1} d\Lambda_k^*(s).$$

Further, the \hat{S}_k^* for different k 's are independent because they are based on independent samples. Also, $n_k/n \rightarrow \pi_k$ as $n \rightarrow \infty$. We therefore obtain

$$\begin{aligned}
\Omega_{rs} &= n \text{Cov}(\hat{L}_r, \hat{L}_s) \\
&\doteq \text{Cov} \left(\int_0^\infty \sum_{v=1}^K \pi_v^{-\frac{1}{2}} \tilde{H}_{rv}(t) \Delta_v(t) dt, \int_0^\infty \sum_{w=1}^K \pi_w^{-\frac{1}{2}} \tilde{H}_{sw}(u) \Delta_w(u) du \right) \\
&= \sum_{k=1}^K \pi_k^{-1} \int_0^\infty \int_0^\infty \tilde{H}_{rk}(t) \tilde{H}_{sk}(u) \text{Cov}(\Delta_k(t), \Delta_k(u)) dt du \\
&\doteq \sum_{k=1}^K \pi_k^{-1} \int_0^\infty \int_0^\infty \tilde{H}_{rk}(t) \tilde{H}_{sk}(u) S_k^*(t) S_k^*(u) \left\{ \int_0^{\min\{t,u\}} G_k^*(y)^{-1} d\Lambda_k^*(y) \right\} dt du \\
&= \sum_{k=1}^K \pi_k^{-1} \int_0^\infty \left\{ \int_y^\infty \tilde{H}_{rk}(t) S_k^*(t) dt \right\} \left\{ \int_y^\infty \tilde{H}_{sk}(u) S_k^*(u) du \right\} G_k^*(y)^{-1} d\Lambda_k^*(y) \\
&= \sum_{k=1}^K \pi_k^{-1} B_{rk} B_{sk} R_{rsk},
\end{aligned}$$

with R_{rsk} given in (5).

The foregoing development parallels that presented by Gill (1983, Cor. 3.2) for the variance of the integrated Kaplan-Meier curve and by Pepe and Fleming (1991, Eqn. 2.2) for the variance of a weighted integral. We note a misprint in Pepe and Fleming's formula: in their notation, their dt should be replaced by $-dS(t)$.

A.3. Asymptotic Theory for the Nonlinear Least Squares Estimator

We denote here the true value of γ by γ^* for emphasis. We assume that the parameter space is compact, that γ^* is an interior point, and that equation $\mathbf{L}(\gamma) = \mathbf{L}(\gamma^*)$ has a unique solution at $\gamma = \gamma^*$. The latter condition will hold if $\mathbf{W}(\gamma)$ has full rank for all γ . We may write the objective function $\Sigma(\gamma)$ as $\Sigma(\gamma) = \{\mathbf{L}(\gamma) - \mathbf{L}(\gamma^*) - \epsilon\}^T \hat{\Omega}^{-1} \{\mathbf{L}(\gamma) - \mathbf{L}(\gamma^*) - \epsilon\}$. Now as $n \rightarrow \infty$, $\epsilon \rightarrow \mathbf{0}$ and $\hat{\Omega} \rightarrow \Omega$ in probability. Thus, the limiting form of the objective function is $\{\mathbf{L}(\gamma) - \mathbf{L}(\gamma^*)\}^T \Omega^{-1} \{\mathbf{L}(\gamma) - \mathbf{L}(\gamma^*)\}$. This function will have a unique minimum at γ^* under the condition that $\mathbf{L}(\gamma) = \mathbf{L}(\gamma^*)$ only at $\gamma = \gamma^*$. It follows that $\hat{\gamma}$ is consistent for γ^* ; cf. the proof of Jennrich (1969, Theorem 6). Since γ^* is an interior point, $\hat{\gamma}$ will satisfy the estimating equation (6) when n is large enough. By Taylor expansion, we may write

$$\mathbf{W}(\hat{\gamma})^T \hat{\Omega}^{-1} \mathbf{W}(\hat{\gamma})(\gamma^* - \hat{\gamma}) + \mathbf{W}(\hat{\gamma})^T \hat{\Omega}^{-1} \epsilon = \mathbf{0}, \quad (18)$$

where $\tilde{\gamma}$ lies between $\hat{\gamma}$ and γ^* . By the consistency of γ and $\hat{\Omega}$, we can write

$$\sqrt{n}(\hat{\gamma} - \gamma^*) \doteq (\mathbf{W}(\gamma^*)^T \Omega^{-1} \mathbf{W}(\gamma^*))^{-1} \mathbf{W}(\gamma^*)^T \Omega^{-1} \{\sqrt{n} \epsilon\}.$$

Because $\sqrt{n} \epsilon$ is asymptotically $N(\mathbf{0}, \Omega)$, we may conclude that $\sqrt{n}(\hat{\gamma} - \gamma^*)$ is asymptotically mean-zero normal with covariance matrix $(\mathbf{W}^T \Omega^{-1} \mathbf{W})^{-1}$.

A.4. Minimizing the Quantity ρ Appearing in the ‘‘Small β ’’ Variance

We recall from (11) the definition of ρ :

$$\rho = \frac{\int_0^\infty \left\{ \int_s^\infty \eta(t) dt \right\}^2 G(s)^{-1} d\Lambda_0(s)}{\left\{ \int_0^\infty \eta(s) \Lambda_0(t) ds \right\}^2}.$$

To minimize ρ , we use an argument of Pepe and Fleming (1991, Sec. 3.3). We define

$$\kappa(s) = \int_s^\infty \eta(t) dt,$$

so that $\kappa'(s) = -\eta(s)$. We then have

$$\rho = \frac{\int_0^\infty \kappa(s)^2 G(s)^{-1} d\Lambda_0(s)}{\left\{ \int_0^\infty \kappa'(s) \Lambda_0(s) ds \right\}^2} = \frac{\int_0^\infty \kappa(s)^2 G(s)^{-1} d\Lambda_0(s)}{\left\{ \int_0^\infty \kappa(s) d\Lambda_0(s) \right\}^2}.$$

By the Cauchy-Schwarz inequality, after inserting $G(s)^{\frac{1}{2}} G(s)^{-\frac{1}{2}}$ on the left hand side,

$$\left\{ \int_0^\infty \kappa(s) d\Lambda_0(s) \right\}^2 \leq \left\{ \int_0^\infty G(s) d\Lambda_0(s) \right\} \left\{ \int_0^\infty G(s)^{-1} \kappa(s)^2 d\Lambda_0(s) \right\}.$$

Thus

$$\rho \geq \left\{ \int_0^\infty G(s) d\Lambda_0(s) \right\}^{-1}, \quad (19)$$

with equality for $\kappa(t) = G(t)$. The optimal $\eta(t)$ is thus given by $\eta(t) dt = -dG(t)$.

REFERENCES

- Andersen, P. K. (1983). Comparing survival distributions via hazard ratio estimates. *Scandinavian Journal of Statistics* **10**, 77-85.
- Armitage, P. and Doll, R. (1961). Stochastic models for carcinogenesis. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability: Biology and Problems of Health*. Berkeley, CA: University of California Press, pp. 19-38.
- Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications*. New York: John Wiley.
- Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2**, 437-453.
- Breslow, N. and Day, N. E. (1993). *Statistical Methods in Cancer Research. Vol. II: The Design and Analysis of Cohort Studies*. Oxford: Oxford University Press.
- Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman-Hall.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.
- Gill, R. D. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Annals of Statistics* **11**, 49-58.
- Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge: Cambridge University Press.
- Hunter, D. J., Rimm, E. B., Sacks, F. M., Stampfer, M. J., Colditz, G. A., Litin, L. B., and Willett, W. C. (1992). Comparison of measures of fatty acid intake by subcutaneous fat aspirate, food frequency questionnaire, and diet records in a free-living population of US men. *American Journal of Epidemiology* **135**, 418-427.

- Hu, F. B., Stampfer, M. J., Manson, J. E., Rimm, E., Colditz, G. A., Rosner, B. A., Hennekens, C. H., and Willett, W. C. (1997). Dietary fat intake and the risk of coronary heart disease in women. *New England Journal of Medicine* **337**, 1491-1499.
- Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics* **40**, 633-643.
- Kalbfleisch, J. and Prentice, R. L. (1981). Estimation of the average hazard ratio. *Biometrika* **68**, 105-112.
- London, S. J., Sacks, F. M., Caesar, J., Stampfer, M. J., Siguel, E., and Willett, W. C. (1991). Fatty acid composition of subcutaneous adipose tissue and diet in post-menopausal US women. *American Journal of Clinical Nutrition* **54**, 340-345.
- Pepe, M. and Fleming, T. (1991). Weighted Kaplan-Meier statistics: large sample and optimality considerations. *Journal of the Royal Statistical Society, Series B* **53**, 341-352.
- Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-342.
- Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York: John Wiley.
- Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. New York: John Wiley.
- Spiegelman, D., Carroll, R. J. and Kipnis, V. (2001). Efficient regression calibration in main study / internal validation study designs with an imperfect reference instrument. *Statistics in Medicine* **20**, 139-160.
- Spiegelman, D., and Gray, R. (1991). Cost-efficient study designs for binary response data with generalized Gaussian measurement error in the covariate. *Biometrics* **47**, 851-870.

- Steinberg, D. M. and Hunter, W. G. (1984). Experimental design: review and comment. *Technometrics* **26**, 71-130.
- Thomas, D. C. (1981). General relative-risk models for survival time and matched case-control analysis. *Biometrics* **37**, 673-686.
- Willett, W. C. (1998). *Nutritional Epidemiology*, 2nd ed. New York: Oxford University Press.
- Willett, W. C., Sampson, C., Stampfer, M. J., Rosner, B. Bain, C., Witschi, J., Hennekens, C. H., and Speizer, F. E. (1985). The reproducibility and validity of the semiquantitative food frequency questionnaire. *American Journal of Epidemiology* **122**, 51-65.
- Zhou, H. and Pepe, M. (1995). Auxilliary covariate data in failure time regression. *Biometrika* **82**, 139-149.
- Zhou, H. and Wang, C. (2000). Failure time regression with continuous covariates measured with error. *Journal of the Royal Statistical Society, Series B* **62**, 657-665.

Table 1

Simulation Results for the Case of a Single Binary Covariate
 Sample Size = 2,000, Unexposed Cumulative Incidence = 25%

Percent Exposed	Error Rate	True RR	Percent Bias In Estimated Log RR			Empirical Variance ($\times 100$)		MSE Ratio	95% CI Coverage
			Naive Cox	WTKME	FWMLE	WTKME	FWMLE		
5 %	1 %	1.5	-15.89	-1.72	-1.92	3.84	3.82	0.99	95.18
5 %	1 %	2.0	-14.00	-0.24	-0.43	3.15	3.10	0.98	95.48
5 %	5 %	1.5	-48.45	-4.48	-4.77	7.00	6.98	1.00	95.94
5 %	5 %	2.0	-46.08	-2.07	-2.16	5.40	5.34	0.99	96.04
5 %	10 %	1.5	-66.02	-6.60	-6.99	12.94	12.53	0.97	96.40
5 %	10 %	2.0	-64.16	-2.60	-2.90	9.89	9.76	0.99	96.44
5 %	20 %	1.5	-82.24	-18.21	-23.81	38.43	42.19	1.11	96.36
5 %	20 %	2.0	-80.92	-6.43	-9.87	29.09	29.15	1.01	97.52
25 %	1 %	1.5	-3.03	-0.02	-0.02	0.93	0.93	1.00	94.90
25 %	1 %	2.0	-3.01	-0.11	-0.10	0.81	0.81	1.00	94.52
25 %	5 %	1.5	-14.09	0.07	0.06	1.13	1.13	1.00	94.72
25 %	5 %	2.0	-13.89	-0.24	-0.22	1.00	0.99	1.00	94.92
25 %	10 %	1.5	-26.61	0.02	0.04	1.50	1.50	1.00	95.06
25 %	10 %	2.0	-26.14	-0.30	-0.30	1.25	1.25	1.00	95.28
25 %	20 %	1.5	-48.41	-0.49	-0.49	2.78	2.78	1.00	95.54
25 %	20 %	2.0	-47.42	-0.16	-0.15	2.45	2.43	0.99	95.10
40 %	1 %	1.5	-2.43	-0.33	-0.32	0.77	0.77	1.00	94.64
40 %	1 %	2.0	-2.11	0.00	0.01	0.66	0.66	1.00	95.20
40 %	5 %	1.5	-10.03	0.42	0.43	0.91	0.91	1.00	94.58
40 %	5 %	2.0	-10.35	0.05	0.07	0.83	0.82	1.00	94.94
40 %	10 %	1.5	-20.63	0.04	0.05	1.16	1.16	1.00	94.76
40 %	10 %	2.0	-20.40	0.31	0.30	1.04	1.03	1.00	95.24
40 %	20 %	1.5	-41.04	-0.32	-0.33	1.97	1.97	1.00	95.76
40 %	20 %	2.0	-40.66	0.31	0.34	1.83	1.83	1.00	95.12

Legend:

RR = Relative Risk

MSE = Mean Square Error

WTKME = Weighted Transformed Kaplan-Meier Estimator

FWMLE = Full Weibull Maximum Likelihood Estimator

Table 2

Simulation Results for the Case of a Single Binary Covariate
 Sample Size = 10,000, Control Cumulative Incidence = 5%

Percent Exposed	Error Rate	True RR	Percent Bias In Estimated Log RR			Empirical Variance ($\times 100$)		MSE Ratio	95% CI Coverage
			Naive Cox	WTKME	FWMLE	WTKME	FWMLE		
5 %	1 %	1.5	-16.70	-3.74	-3.74	3.70	3.70	1.00	95.28
5 %	1 %	2.0	-13.21	-0.98	-0.97	2.69	2.69	1.00	95.36
5 %	5 %	1.5	-47.94	-6.87	-6.86	6.35	6.35	1.00	95.94
5 %	5 %	2.0	-43.53	-2.55	-2.54	4.44	4.44	1.00	95.10
5 %	10 %	1.5	-65.01	-9.15	-9.18	11.20	11.20	1.00	95.94
5 %	10 %	2.0	-61.59	-3.11	-3.65	7.00	6.84	0.98	96.02
5 %	20 %	1.5	-81.70	-24.44	-38.33	36.33	44.45	1.26	94.89
5 %	20 %	2.0	-79.51	-9.67	-17.01	20.62	21.85	1.10	95.77
25 %	1 %	1.5	-3.76	-0.92	-0.92	0.87	0.87	1.00	94.86
25 %	1 %	2.0	-2.93	-0.27	-0.27	0.72	0.72	1.00	95.42
25 %	5 %	1.5	-14.24	-0.70	-0.70	1.04	1.04	1.00	95.08
25 %	5 %	2.0	-13.09	-0.31	-0.31	0.87	0.87	1.00	95.22
25 %	10 %	1.5	-26.60	-1.00	-1.00	1.36	1.36	1.00	95.40
25 %	10 %	2.0	-24.57	-0.01	0.00	1.15	1.15	1.00	95.06
25 %	20 %	1.5	-47.77	-1.12	-1.10	2.56	2.56	1.00	95.40
25 %	20 %	2.0	-45.79	-0.13	-0.12	2.04	2.05	1.00	95.44
40 %	1 %	1.5	-2.44	-0.38	-0.38	0.74	0.74	1.00	94.92
40 %	1 %	2.0	-2.14	-0.09	-0.09	0.64	0.64	1.00	95.44
40 %	5 %	1.5	-10.28	-0.02	-0.02	0.84	0.84	1.00	95.60
40 %	5 %	2.0	-9.93	0.32	0.33	0.74	0.74	1.00	95.42
40 %	10 %	1.5	-20.84	-0.51	-0.50	1.12	1.12	1.00	95.08
40 %	10 %	2.0	-20.29	0.10	0.10	0.97	0.97	1.00	95.00
40 %	20 %	1.5	-40.48	0.16	0.16	2.00	2.00	1.00	95.26
40 %	20 %	2.0	-40.63	-0.19	-0.18	1.78	1.78	1.00	94.86

Legend:

RR = Relative Risk

MSE = Mean Square Error

WTKME = Weighted Transformed Kaplan-Meier Estimator

FWMLE = Full Weibull Maximum Likelihood Estimator

Table 3
 Estimated Coefficients and Standard Errors for the Nurses Health Study of the
 Relationship Between TFA Intake and Cardiovascular Disease

Method	TFA		AGE2		AGE3		AGE4	
	Estimate	Std Err	Estimate	Std Err	Estimate	Std Err	Estimate	Std Err
Cox	0.2616	0.0771	1.0668	0.1101	1.6724	0.1004	1.8768	0.1019
I-NI	0.2731	0.0772	1.0662	0.1101	1.6708	0.1004	1.8777	0.1020
I-I	0.2415	0.0780	1.0717	0.1094	1.6800	0.0997	1.8651	0.1018
A0-NI	1.0690	0.3649	1.0319	0.1116	1.6130	0.1083	2.0216	0.2020
A0-I	1.0706	0.3052	1.0726	0.1094	1.6825	0.0998	1.8585	0.1034
A1-NI	0.5438	0.4851	1.0586	0.1130	1.6741	0.1157	2.2810	0.2657
A1-I	0.4915	0.4553	1.0648	0.1099	1.6673	0.1024	1.8284	0.1420

Legend:

TFA = indicator variable for high TFA intake
 AGE2 = indicator variable for 2nd age stratum (45-49)
 AGE3 = indicator variable for 3rd age stratum (50-54)
 AGE4 = indicator variable for 4th age stratum (55 +)

Cox = classical Cox regression analysis
 I-NI = our method, identity \mathbf{A} matrix, non-iterative estimator
 I-I = our method, identity \mathbf{A} matrix, iterative estimator
 A0-NI = our method, observed \mathbf{A} matrix taken as known, non-iterative estimator
 A0-I = our method, observed \mathbf{A} matrix taken as known, iterative estimator
 A1-NI = our method, accounting for uncertainty in \mathbf{A} matrix, non-iterative estimator
 A1-I = our method, accounting for uncertainty in \mathbf{A} matrix, iterative estimator