Estimating Predictors for Long- Or Short-Term Survivors

Lu Tian*

Wei Wang[†]

L. J. Wei[‡]

*Harvard School of Public Health, ltian@hsph.harvard.edu

[†]Harvard University Biostatistics, wwang@hsph.harvard.edu

[‡]Harvard School of Public Health, wei@hsph.harvard.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/harvardbiostat/paper8>

Copyright ©2003 by the authors.

ESTIMATING PREDICTORS FOR LONG- OR SHORT-TERM SURVIVORS

L. Tian, W. Wang and L.J. Wei

Department of Biostatistics, Harvard School of Public Health

677 Huntington Ave., Boston, MA 02115

email: ltian@hsph.harvard.edu, wwang@hsph.harvard.edu, wei@hsph.harvard.edu

SUMMARY. Suppose that the response variable in a well-executed clinical or observational study to evaluate a treatment is the time to a certain event, and a set of baseline covariates or predictors was collected for each study patient. Furthermore, suppose that a significant number of study patients had non-trivial, long-term adverse effects from the treatment. A commonly posed question is how to use these covariates from the study to identify future patients who would (or would not) benefit from the treatment. In this article, we present “point” and “interval” estimates for the *set* of covariate or predictor vectors associated with a specific patient survival status, e.g., long (or short) term survival, in the presence of censoring. These estimates can be easily displayed on a two-dimensional plane even for the case with high dimensional covariate vectors. These simple numerical and graphical procedures provide useful information for patient management and/or the design of future studies, which are key issues in pharmacogenomics with genetic markers. The new proposal is illustrated with a data set from a cancer study for treating multiple myeloma.

KEY WORDS: Breslow estimate; Calibration; Censoring; Cox proportional hazards model; Pharmacogenomics; Prediction.

1. INTRODUCTION

Suppose that a clinical trial or observational study whose efficacy endpoint is the time to a specific event was successfully carried out for evaluating a treatment. The application of the treatment to the general patient population, however, is not transparent due to, for example, its non-trivial, long-term adverse effects. Suppose that for each study patient, a set of potentially important covariates including certain bio- and genetic-markers was collected at the patient's entry time. An interesting and important question is how to use these predictors to identify future subjects who would (or would not) be good candidates for receiving the treatment. For example, in a clinical study conducted by the Eastern Cooperative Oncology Group (ECOG) to explore whether adding interferon (IFN) to the standard chemotherapy (VBMCP) for treating multiple myeloma would increase the patient's survival, more than twenty-nine baseline covariates were recorded for each subject (Greipp et al. (1998); Oken et al. (1999)). With 239 patients treated by this specific combination therapy, on average there is no obvious extra benefit with respect to patient survival, but IFN + VBMCP does yield a higher complete response rate and longer response duration than VBMCP. The median survival time for study patients is about 42 months. Since the treatment period for patients with IFN is rather long (up to two years) and IFN has some serious, persistent adverse effects including acute flu syndrome, chronic fatigue, among others, it does not seem desirable to treat future myeloma patients with IFN, who would die within, for example, two years, or to recruit such patients for new studies on IFN. On the other hand, the benefit of adding IFN to the standard chemotherapy may well outweigh the toxicity for patients who would survive, say, more than five years. Therefore, it is important to know how to utilize those baseline covariates collected in the study to identify potential long- or short-term survivors to be treated by IFN + VBMCP.

To tackle this problem in the presence of censored failure times, one usually fits the data with a Cox proportional hazards model, which relates the hazard function of the individual event time to its covariates (Cox (1972)). Specifically, let T be the failure time and $\mathbf{Z}$ be the corresponding $d \times 1$ random vector of covariates or predictors. Assume that $\lambda_{\mathbf{z}}(t)$, the hazard function of T given $\mathbf{Z} = \mathbf{z}$ at time t , is related to $\mathbf{z}$ via the standard Cox proportional hazards model

$$\lambda_{\mathbf{z}}(t) = \lambda_0(t)e^{\beta_0' \mathbf{z}}, \quad (1.1)$$

where $\lambda_0(t)$ is a completely unspecified function of t and β_0 is the vector of the unknown regression parameters. After fitting the data with the Cox model (1.1), suppose that we are interested in identifying *all* the future patients based on their covariate vectors $\mathbf{z}$, who “likely” survive beyond t_0 , a relatively large time point. For instance, one may consider the following set of covariate vectors:

$$\Omega_p = \{\mathbf{z} : S_{\mathbf{z}}(t_0) \geq p\}, \quad (1.2)$$

where $0 < p < 1$ and $S_{\mathbf{z}}(t)$ is the survival probability at time t for patients with the covariate vector $\mathbf{z}$ (e.g., $t_0 = 5$ years and $p = 0.8$). Patients whose covariate vector $\mathbf{z} \in \Omega_p$ are expected to be long-term survivors. A reasonable “point” estimator for Ω_p is

$$\hat{\Omega}_p = \{\mathbf{z} : \hat{S}_{\mathbf{z}}(t_0) \geq p\}, \quad (1.3)$$

where $\hat{S}_{\mathbf{z}}(t)$ is obtained via the Breslow estimate for $\Lambda_{\mathbf{z}}(t)$, the cumulative hazard function of $\lambda_{\mathbf{z}}(t)$ in (1.1) (Breslow (1972); Altman and Andersen (1986); Lin, Fleming and Wei (1994)). In Section 2 of this paper, we show that this estimator is strongly consistent for Ω_p .

Although the dimension of the covariate vector may be large, the observed estimate $\hat{\Omega}_p$ can be determined by the patient’s estimated risk score $\hat{\beta}' \mathbf{z}$, where $\hat{\beta}$ is the maximum partial likelihood estimate for β_0 . This univariate score, which combines the information

across all the covariates of a patient, has been used extensively in predicting breast cancer risk (Gail et al. (1989)). To make further inferences about Ω_p , one may construct a $(1 - \alpha)$ confidence “interval” $(\underline{\Omega}_p, \overline{\Omega}_p)$ for this parameter (which is a subset of R^d) such that $\Pr(\underline{\Omega}_p \subset \Omega_p \subset \overline{\Omega}_p) \approx 1 - \alpha$, where $0 < \alpha < 1$. Patients whose covariate vectors fall into $\underline{\Omega}_p$ are likely to live long. On the other hand, patients would less likely be long-term survivors if their covariate vectors are in the complement of $\overline{\Omega}_p$. Now, let us consider a natural candidate for this type of interval estimator whose lower bound $\underline{\Omega}_p$ is

$$\{\mathbf{z} : \widehat{S}_{\mathbf{z}}(t_0) - \eta_{\alpha/2} \widehat{\tau}_{\mathbf{z}} \geq p\}, \quad (1.4)$$

and whose upper bound is the complement of

$$\{\mathbf{z} : \widehat{S}_{\mathbf{z}}(t_0) + \eta_{\alpha/2} \widehat{\tau}_{\mathbf{z}} < p\}, \quad (1.5)$$

where $\widehat{\tau}_{\mathbf{z}}$ is an estimated standard error of $\widehat{S}_{\mathbf{z}}(t_0)$, and $\eta_{\alpha/2}$ is the upper $100(\alpha/2)$ percentage point of the standard normal. That is, for a given $\mathbf{z}$, if the lower bound of the standard confidence interval for $S_{\mathbf{z}}(t_0)$ is larger than p , we label this patient as a long term survivor. On the other hand, if the upper bound of the interval is below p , this patient is not expected to live long, and we assign this subject to the complement of $\overline{\Omega}_p$. Asymptotically this interval has the correct coverage probability for Ω_p with a *single* prediction. This simple procedure, however, is not valid asymptotically for multiple predictions with two or more distinct covariate vectors $\{\mathbf{z}\}$ such that $S_{\mathbf{z}}(t_0) = p$. Moreover, for practical sample sizes, even the covariate vectors are not on the boundary of Ω_p , the above procedure with bounds (1.4) and (1.5) may give us a rather distorted picture of Ω_p . For instance, in one case considered in Section 3, we show via simulation that the empirical probability that $(1.4) \subset \Omega_p \subset (1.5)$ is less than 0.66 with $1 - \alpha = 0.95$.

In this paper, we present the confidence bounds $\underline{\Omega}_p$ and $\overline{\Omega}_p$ for Ω_p , which are valid for the case with any number of predictions. These two bounds can be readily displayed on

a two-dimensional plane even for the case with high dimensional covariate vectors. This type of interval estimation is quite useful for guiding us in improving patient care and designing future studies for similar regimens. We illustrate the new proposal using a data set from the study on treating multiple myeloma with IFN + VBMCP. Conceptually our proposal is closely related to the procedures for solving the classical calibration problem with multivariate predictors and non-censored observations (Lee (1998)).

2. INTERVAL ESTIMATES FOR A SET OF COVARIATE VECTORS FOR LONG- OR SHORT-TERM SURVIVORS

Let C be the censoring variable for T . Conditional on $\mathbf{Z}$, T and C are assumed to be independent of each other. For T , one can only observe (X, Δ) , where $X = \min\{T, C\}$, $\Delta = I(X = T)$, and $I(\cdot)$ is the indicator function. Our data consist of n independent and identical copies, $\{(X_i, \mathbf{Z}_i, \Delta_i), i = 1, \dots, n\}$, of $(X, \mathbf{Z}, \Delta)$. The estimator $\hat{\Omega}_p$ in (1.3) is “consistent” for the set Ω_p . Specifically, we define the distance between two sets A and B in R^d by $d(A, B) = \text{pr}_{\mathbf{Z}}(A \cup B - A \cap B)$, where $\text{pr}_{\mathbf{Z}}(\cdot)$ is the probability on R^d generated by the covariate vector $\mathbf{Z}$. If $\mathbf{Z}$ has a bounded support, it is straightforward to show that

$$d(\hat{\Omega}_p, \Omega_p) \leq \text{constant} \times (\|\hat{\beta} - \beta_0\| + |\hat{S}_0(t_0) - S_0(t_0)|) \rightarrow 0,$$

almost surely. That is, for a large n , we expect that $\hat{\Omega}_p$ is almost overlapped with Ω_p .

To obtain the bounds for Ω_p , we may consider a similar type of bounds (1.4) and (1.5) with an appropriate adjustment on the cutoff point η_α . Since $S_{\mathbf{z}}(t_0)$ is bounded between 0 and 1, we consider a complementary log-log transformation of this probability. To this end, we rewrite the estimate $\hat{\Omega}_p$ as

$$\{\mathbf{z} : \log(-\log(\hat{S}_0(t_0))) + \hat{\beta}'\mathbf{z} \leq \log(-\log p)\}.$$

To obtain confidence “bounds” for Ω_p , let us consider the process of $\mathbf{z}$

$$W(\mathbf{z}) = n^{1/2} \left(\log(-\log(\hat{S}_0(t_0))) + \hat{\beta}'\mathbf{z} - \log(-\log(S_0(t_0))) - \beta_0'\mathbf{z} \right).$$

It is straightforward to show that this process converges weakly to a Gaussian process in $\mathbf{z}$. Now, let $\sigma(\mathbf{z})$ be an estimated standard error of $W(\mathbf{z})$, and let $c_\alpha \geq 0, 0 < \alpha < 1$, be a cut-off point such that,

$$\text{pr} \left(\sup_{\mathbf{z} \in \mathcal{H}} \{W(\mathbf{z})/\sigma(\mathbf{z})\} \leq c_\alpha \right) \approx \text{pr} \left(\inf_{\mathbf{z} \in \mathcal{H}} \{W(\mathbf{z})/\sigma(\mathbf{z})\} \geq -c_\alpha \right) \approx 1 - \alpha, \quad (2.1)$$

where $\mathcal{H}$ is the set of $\mathbf{z}$ which we are interested in making prediction from. Define a set

$$\underline{\Omega}_p = \{\mathbf{z} : \log(-\log(\widehat{S}_0(t_0))) + \widehat{\boldsymbol{\beta}}' \mathbf{z} + n^{-1/2} c_\alpha \sigma(\mathbf{z}) \leq \log(-\log p)\}. \quad (2.2)$$

Then it follows from (2.1) that

$$\text{pr} \left(\log(-\log(\widehat{S}_0(t_0))) + \widehat{\boldsymbol{\beta}}' \mathbf{z} - \log(-\log(S_0(t_0))) - \boldsymbol{\beta}'_0 \mathbf{z} \geq -n^{-1/2} c_\alpha \sigma(\mathbf{z}), \forall \mathbf{z} \in \mathcal{H} \right) \approx 1 - \alpha.$$

This implies that

$$\text{pr} (\underline{\Omega}_p \subset \Omega_p, 0 < p < 1) =$$

$$\text{pr} \left(\log(-\log(\widehat{S}_0(t_0))) + \widehat{\boldsymbol{\beta}}' \mathbf{z} + n^{-1/2} c_\alpha \sigma(\mathbf{z}) \geq \log(-\log(S_0(t_0))) + \boldsymbol{\beta}'_0 \mathbf{z}, \forall \mathbf{z} \in \mathcal{H} \right),$$

and

$$\text{pr}(\underline{\Omega}_p \subset \Omega_p, 0 < p < 1) \approx 1 - \alpha. \quad (2.3)$$

Note that the probability statement in (2.3) is valid uniformly in p , and c_α is heavily dependent on the choice of $\mathcal{H}$. For instance, one may let $\mathcal{H}$ be a specific subset of the support of the covariate random vector $\mathbf{Z}$.

To approximate the cutoff point c_α in (2.1), consider the random vector

$$n^{1/2} \left(\log(-\log(\widehat{S}_0(t_0))) - \log(-\log(S_0(t_0))), \widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}'_0 \right)'.$$

It follows from Anderson and Gill (1982) and Lin et al. (1994) that this vector converges in distribution to a multivariate normal, say, G with mean 0. The covariance matrix of

G can be consistently estimated by $\Sigma = \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & n\mathcal{I} \end{pmatrix}$, where

$$\sigma_{11} = n\hat{\Lambda}^{-2}(t_0) \left[\int_0^{t_0} \frac{dN(s)}{S^{(0)}(\hat{\boldsymbol{\beta}}, s)^2} + \left\{ \int_0^{t_0} \frac{S^{(1)}(\hat{\boldsymbol{\beta}}, s)' dN(s)}{S^{(0)}(\hat{\boldsymbol{\beta}}, s)^2} \right\} \mathcal{I} \left\{ \int_0^{t_0} \frac{S^{(1)}(\hat{\boldsymbol{\beta}}, s) dN(s)}{S^{(0)}(\hat{\boldsymbol{\beta}}, s)^2} \right\} \right],$$

$$\boldsymbol{\sigma}_{21} = \boldsymbol{\sigma}_{12}' = -n\hat{\Lambda}^{-1}(t_0) \mathcal{I} \left\{ \int_0^{t_0} \frac{S^{(1)}(\hat{\boldsymbol{\beta}}, s) dN(s)}{S^{(0)}(\hat{\boldsymbol{\beta}}, s)^2} \right\},$$

$$S^{(r)}(\boldsymbol{\beta}, s) = \sum_{i=1}^n I(X_i \geq s) \mathbf{Z}_i^r e^{\boldsymbol{\beta}' \mathbf{Z}_i}, r = 0, 1,$$

$$N(s) = \sum_{i=1}^n I(X_i \leq s, \Delta_i = 1),$$

$\mathcal{I}$ is the inverse of the observed information matrix of the partial likelihood function for model (2.1). Thus, the process $W(\mathbf{z})/\sigma(\mathbf{z})$ converges in distribution to $\frac{(1, \mathbf{z}')G}{\{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'\}^{1/2}}$, as $n \rightarrow \infty$.

If $\mathcal{H}$ is the support of $\mathbf{Z}$ and if the component of $\mathbf{Z}$ is either discrete (with a finite number of possible values) or continuous, $\mathcal{H}$ can be defined by a set of linear constraints of $\mathbf{z}$. For any realization of G , $\sup_{\mathbf{z} \in \mathcal{H}} (1, \mathbf{z}')G / \{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'\}^{1/2}$ can be obtained via the standard quadratic programming technique. Approximations to the distribution of $\sup\{W(\mathbf{z})/\sigma(\mathbf{z})\}$ and the corresponding c_α can then be obtained by simulating a large number of realizations from G (see Appendix for details). If we let $\mathcal{H} = R^d$, the distribution of such a sup-statistic can be approximated by the distribution of a mixture of two random variables: $\frac{1}{2}\chi_d + \frac{1}{2}\chi_{d+1}$. One may use the $100(1 - \alpha)$ th percentile of this mixture as a conservative approximation to c_α in (2.1). The detail of the above approximation is also given in the Appendix.

Similarly, to obtain an outer “bound” of Ω_p , define the set

$$\overline{\Omega}_p = \{\mathbf{z} : \log(-\log(\hat{S}_0(t_0))) + \hat{\boldsymbol{\beta}}' \mathbf{z} - n^{-1/2} c_\alpha \sigma(\mathbf{z}) \leq \log(-\log p)\}. \quad (2.4)$$

It follows that $\text{pr}(\Omega_p \subset \overline{\Omega}_p, 0 < p < 1) \approx 1 - \alpha$. Therefore, a patient whose $\mathbf{z}$ does not belong to $\overline{\Omega}_p$ would be less likely to survive beyond t_0 .

It is important to note that although the observed sets $\underline{\Omega}_p$ and $\overline{\Omega}_p$ are functions of a possibly high dimensional covariate vector $\mathbf{z}$, they are uniquely determined by $x = \sigma(\mathbf{z})$ and the risk score $y = \hat{\beta}'\mathbf{z}$. Therefore, the above two sets can be simply displayed on the (x, y) -plane. In Figure 1, we present a typical plot for these sets. First, the set of the point estimate $\hat{\Omega}_p$ in (1.2) is the entire region under the horizontal line $y = \hat{\beta}'\mathbf{z} = \log(-\log p) - \log(-\log(\hat{S}_0(t_0)))$. The heavily shaded area is $\underline{\Omega}_p$ and the lightly shaded area is the complement of $\overline{\Omega}_p$. The straight lines, which define the boundaries of these two sets, depend on the confidence level α . This graphical display is quite informative. For example, the black dot in the plot is closer to the above horizontal line than the open circle point, indicating that we are much more confident that the patient associated with the open circle point is a long-term survivor than the patient corresponding to the black dot. Moreover, although the subject with “*” has a larger risk score $\hat{\beta}'\mathbf{z}$ than that with the black dot, this subject is likely to be a long term survivor due to its small $\sigma(\mathbf{z})$.

One may consider an alternative way to define a long-term survivor. For instance, consider

$$\Theta_\tau = \{\mathbf{z} : \tau^{th}(T) \geq t_0\}, \quad (2.5)$$

where $\tau^{th}(T)$ is the τ^{th} quantile of T conditional on $\mathbf{z}$. For example, one may be interested in Θ_τ with $\tau = 1/2$ and $t_0 = 5$ years. Note that Θ_τ is equivalent to $\{\mathbf{z} : S_0(t_0)^{\exp(\beta'_0\mathbf{z})} \geq 1 - \tau\}$. The corresponding two bounds $\underline{\Theta}_\tau$ and $\overline{\Theta}_\tau$ for (2.5) can be obtained by replacing p in (2.2) and (2.4) with $(1 - \tau)$, respectively.

One may also be interested in identifying future patients based on $\mathbf{z}$, who tend to die relatively early. For this case, let us consider

$$\Omega_p^* = \{\mathbf{z} : \text{pr}(T < t_0 | \mathbf{z}) \geq p\} = \{\mathbf{z} : S_0(t_0)^{\exp(\beta'_0\mathbf{z})} \leq 1 - p\}, \quad (2.6)$$

where t_0 is a relatively small time point. The corresponding two bounds for Ω_p^* are

$$\begin{aligned}\underline{\Omega}_p^* &= \{\mathbf{z} : \log(-\log(\hat{S}_0(t_0))) + \hat{\beta}'\mathbf{z} - n^{-1/2}c_\alpha\sigma(\mathbf{z}) \geq \log(-\log(1-p))\}, \\ \overline{\Omega}_p^* &= \{\mathbf{z} : \log(-\log(\hat{S}_0(t_0))) + \hat{\beta}'\mathbf{z} + n^{-1/2}c_\alpha\sigma(\mathbf{z}) \geq \log(-\log(1-p))\}.\end{aligned}\quad (2.7)$$

Similar to (2.5), one may use a quantile of T to define short-term survivors by defining

$$\Theta_\tau^* = \{\mathbf{z} : \tau^{th}(T) \leq t_0\} = \{\mathbf{z} : S_0(t_0)^{\exp(\beta'_0\mathbf{z})} \leq 1 - \tau\}.\quad (2.8)$$

The bounds $\underline{\Theta}_\tau^*$ and $\overline{\Theta}_\tau^*$ for (2.8) can be obtained by replacing p in (2.7) with $1 - \tau$ accordingly.

Now, suppose that we are interested in obtaining $\underline{\Omega}_p$ and $\overline{\Omega}_p$, which are valid only for the case with a single p (not uniformly in p). That is,

$$\text{pr}(\underline{\Omega}_p \subset \Omega_p) = \text{pr}(\Omega_p \subset \overline{\Omega}_p) \geq 1 - \alpha.\quad (2.9)$$

For this case, the cutoff point c_α for $\underline{\Omega}_p$ can be obtained via

$$\text{pr}\left(\sup_{\mathbf{z} \in \mathcal{H} \cap \tilde{\Omega}_p^c} (1, \mathbf{z}')G / \{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'\}^{1/2} \geq c_\alpha\right) = \alpha,\quad (2.10)$$

where $\tilde{\Omega}_p^c$ is the complement of the observed value of $\hat{\Omega}_p$. The cutoff point c_α for $\overline{\Omega}_p$ can be obtained via

$$\text{pr}\left(\sup_{\mathbf{z} \in \mathcal{H} \cap \tilde{\Omega}_p} (1, \mathbf{z}')G / \{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'\}^{1/2} \geq c_\alpha\right) = \alpha,\quad (2.11)$$

where $\tilde{\Omega}_p$ is the observed $\hat{\Omega}_p$. Note that these two critical values c_α may be different. Again, numerically these two cutoff points can be obtained via the quadratic programming and the aforementioned simulation technique.

3. EXAMPLE

We illustrate the new numerical and graphical procedures using a data set from the clinical study mentioned in the Introduction on treating patients with multiple myeloma

(MM). Multiple myeloma is a neoplastic disease of bone marrow plasma cells causing lytic bone lesions, anemia, renal insufficiency, and hypercalcemia. The study was designed to evaluate the effect of adding alternating cycles of IFN or early intensification with high dose HiCy to the standard chemotherapy VBMCP regimen for treating MM. Six hundred and twenty-eight patients entered the study from February 1988 to May 1992, and they were randomized to one of the three arms: VBMCP, VBMCP + rIFN $_{\alpha 2}$, and VBMCP + HiCy (Oken et al. (1999)). With respect to overall survival, there are no differences among these three treatment groups (see Figure 2). On the other hand the superior ability of VBMCP + IFN induction therapy to produce complete response and more durable responses suggests that this combination has promising biologic activities in myeloma.

Numerous baseline covariates for each patient were collected for the study and special efforts were made to build a prediction model for MM patient survival (see Greipp et al. (1998)). With the data from all the eligible patients for the study, seven covariates were selected as the important predictors via the Cox regression model for patient survival. These covariates are PB morphology type (PB), plasma cell labeling index (PCLI), $\beta 2$ microglobulin ($\beta 2M$), C-reactive protein (CRP), soluble interleukin-6 receptor (sIL6R), plasma cell percentage by immunofluorescence (PC) and serum creatinine (SC). Note that only the first covariate is discrete (dichotomized). The vector $\hat{\beta}$ of the regression coefficient estimates is $(0.642, 0.14, 0.028, 0.07, 0.912, 0.05, 0.001)'$, and the corresponding estimated standard errors are 0.184, 0.07, 0.012, 0.022, 0.253, 0.019 and 0.0004, respectively. There are 39 patients with plasmablastic MM (PB), and the corresponding covariate value is one for the PB case, 0, otherwise. The PB patients tend to have much shorter survival. The above estimates were obtained based on 447 patients from the entire study cohort, who had complete information on these seven covariates.

From Figure 2, on average, 30% of the study patients died within 2 years and 30% of

them survived beyond 5 years. Here, we present four scenarios to illustrate our proposal with the above fitted Cox model. For the first case, we identify MM patients who are expected to die relatively early. Specifically, consider Ω_p^* in (2.6) with $p = 0.7$ and $t_0 = 42$ months, which is the observed median survival time for the entire cohort. That is, those patients in Ω_p^* are likely to die within 42 months after treatment. The heavily shaded area in Figure 3(a) is the observed $\underline{\Omega}_p^*$ with $\alpha = 0.1$ and $c_\alpha = 3.30$ which was obtained via (2.10) with this specific $p = 0.7$ and $\mathcal{H} = \{0, 1\} \cup R^6$. Note that if we use the cutoff point in (2.1), which is uniform in p , $c_\alpha = 3.45$. Furthermore, if we let $\mathcal{H} = R^7$, the cutoff point is 3.57. The lightly shaded region is the complement of $\overline{\Omega}_p$. Naturally it is interesting to see where the aforementioned 447 patients in the study are located on this graphical display. To this end, each black dot in the figure represents a study patient's estimated risk score $y = \hat{\beta}'\mathbf{z}$ and $x = \sigma(\mathbf{z})$. There are 42 patients in $\hat{\Omega}_{0.7}^*$, the region above the solid horizontal line $y = \log(-\log p) - \log(-\log(\hat{S}_0(t_0))) = 1.6$. On the other hand, there are only 5 study patients in $\underline{\Omega}_p^*$, indicating a high degree of uncertainty of the “point” estimate $\hat{\Omega}_{0.7}^*$. Note that a point in the dark region, which is far away from the solid line $y = 1.6$, represents a short-term survivor with a much higher confidence level than that of a point near the line.

For the second case, we considered Θ_τ^* with $t_0 = 36$ months, a relatively small time point, and $\tau = 0.5$. The darker area in Figure 3(b) is $\underline{\Theta}_{0.5}^*$, with $\alpha = 0.1$. The lightly shaded area corresponds to the complement of $\overline{\Theta}_{0.5}^*$. There are 18 study patients belonging to $\underline{\Theta}_{0.5}^*$, and eleven are PB MM (only 8% of the entire study cohort are PB). These eleven study patients had rather high SC and PC, and had extremely high $\beta 2\text{M}$ and PCLI. On the other hand, the distributions of the other two covariates CRP and sIL6R for these 18 patients are similar to those of the entire cohort.

Next, we considered the case associated with long term survivors by letting $t_0 = 42$

months and $p = 0.7$ in Ω_p . The $\underline{\Omega}_{0.7}$ with $\alpha = 0.10$, is the bottom, darker region in Figure 4(a). In Figure 4(b), we present the case with $t_0 = 48$ months and $\tau = 0.5$ for Θ_τ . There are 52 study patients in the dark region $\underline{\Theta}_{0.5}$. None of these patients is PB and they all had rather low $\beta 2M$, PCLI and PC values.

4. PERFORMANCE OF THE INTERVAL ESTIMATOR FOR SMALL- AND MODERATE-SIZED SAMPLES

A limited simulation study was conducted to examine if the large sample approximation to the distribution of $\sup_{\mathbf{z} \in \mathcal{H}} W(\mathbf{z})/\sigma(\mathbf{z})$ is adequate for practical sample sizes or equivalently, to examine the validity of the confidence bounds in (2.3). To this end, with a two-parameter Weibull regression model, we fitted the data from those 447 patients in the myeloma study, who had complete information on the seven covariates discussed in the Example Section. This results in the Weibull model:

$$\lambda_{\mathbf{z}}(t) = 1.32t^{0.32} \exp(-6.24 + 0.61 * \text{PB} + 0.14 * \text{SC} + 0.03 * \beta 2M + 0.07 * \text{PCLI} + 0.90 * \text{PC} + 0.05 * \text{CRP} + 0.0008 * \text{sIL6R}). \quad (4.1)$$

In our simulation, we considered the case that $\mathcal{H} = \{0, 1\} \times R^6$ for obtaining c_α (therefore, asymptotically the results should be valid uniformly in p). For each fixed sample size n , we repeatedly used Model (4.1) to generate a set of n survival times with the covariate vectors randomly selected from 447 observed covariate vectors with replacement, and the corresponding censoring variable was assumed to be a uniform with support $(0, \tau)$, where τ was chosen to have a pre-specified censoring proportion. In Table 1, we report the empirical coverage probabilities of $\underline{\Omega}_p$ for Ω_p , for $0 < p < 1$, where $t_0 = 42$. Note that for the present case, $\text{pr}(\underline{\Omega}_p \subset \Omega_p, 0 < p < 1) = \text{pr}(\sup_{\mathbf{z} \in \mathcal{H}} \{W(\mathbf{z})/\sigma(\mathbf{z})\} \leq c_\alpha)$, to obtain the empirical coverage levels via simulation, for each generated sample, we check whether the realized $\sup_{\mathbf{z} \in \mathcal{H}} \{W(\mathbf{z})/\sigma(\mathbf{z})\}$ is greater than c_α or not. Moreover, since $W(\mathbf{z})$ is a

linear combination of $\mathbf{z}$ and $\sigma^2(\mathbf{z})$ is a quadratic form of $\mathbf{z}$, $\sup_{\mathbf{z} \in \mathcal{H}} \{W(\mathbf{z})/\sigma(\mathbf{z})\}$ can be obtained analytically via Formula 1f.1.3 of Rao (1973) (p.60). Each entry in the table was obtained based on 10,000 repetitions. The empirical levels of this interval estimator are quite close to their nominal counterparts with moderate sample sizes. On the other hand, when n is less than 100, the coverage probabilities are noticeably less than their nominal counterparts.

Empirically we find that it is possible that a subject, who is classified to be a long-term survivor with respect to t_0 , may fall into the gray zone with respect to a time point earlier than t_0 . This non-monotonicity feature is due to the fact that the estimated standard error of the survival probability estimator at t_0 is smaller than that at that earlier time point.

Lastly, to examine the adequacy of the simple interval estimation procedure with bounds (1.4) and (1.5) discussed in the Introduction, we used Model (4.1) to generate 2000 sets of samples based on the observed covariate vectors from the MM study with $n = 447$. For each realized sample, we obtained the lower and upper bound (1.4) and (1.5) with $\eta_\alpha = 1.96$ for $\Omega_p = \{\mathbf{z} : S_{\mathbf{z}}(42) \geq 0.5\}$, which is equivalent to the set of $\mathbf{z}$ associated with patients whose median survival times are no less than 42 months. Assume that we are only interested in predicting patients with $\mathbf{z}$ in the set of the observed covariate vectors from the MM study. Without censoring involved in these 2000 simulated samples, the empirical probability that $(1.4) \subset \Omega_{0.5} \subset (1.5)$ is only 0.66 instead of 0.95, the nominal level of the interval.

5. REMARKS

The Cox model (1.1) for modeling the hazard function with its covariate can be replaced by other parametric or semi-parametric survival models, for example, the accelerated failure time (Wei (1992)) or linear transformation models (Cheng, Wei and Ying (1995)).

In theory, one may use a non-parametric regression function estimate without modeling the distribution of the failure time with its covariates to estimate $\text{pr}(T > t_0|\mathbf{z})$ to obtain $\hat{\Omega}_p$ or modeling the median of T given $\mathbf{z}$ to construct $\hat{\Theta}_{1/2}$. In practice, especially for the high-dimensional case with censored observations, however, it is rather difficult to implement such a purely nonparametric procedure.

Now, consider the case that there are no censored observations and the response variable T is either continuous or discrete. Furthermore, conditional on $\mathbf{z}$, let the mean of the response variable T be $\mu(\beta'_0\mathbf{z})$, where $\mu(\cdot)$ is a known, strictly increasing function. Then, one may be interested in estimating the set

$$\Omega_a = \{\mathbf{z} : \mu(\beta'_0\mathbf{z}) > a\} = \{\mathbf{z} : \beta'_0\mathbf{z} > \mu^{-1}(a)\},$$

where “a” is a given constant. Let $\hat{\beta}$ be the estimate for β_0 and $\sigma(\mathbf{z})$ be its estimated standard error. The two bounds $\underline{\Omega}_a$ and $\overline{\Omega}_a$ for Ω_a can be constructed and displayed with $y = \hat{\beta}'\mathbf{z}$ and $x = \sigma(\mathbf{z})$ accordingly as in Section 2.

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor, two referees, and Drs. Daniel Commenges, Samuel Heft, Jason Fine and Meredith Goldwasser for insightful and helpful comments on the paper. The work is partially supported by the US NIH grants AI24643 and AI52817.

REFERENCES

- Altman, D. G. and Andersen, P. K. (1986). A note on the uncertainty of a survival probability estimated from Cox’s regression model. *Biometrika* **73**, 722–724.

- Anderson, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10**, 1100–1120.
- Breslow, N. (1972). Contribution to the discussion of "Regression models and life-tables" by Cox, D. R. *Journal of the Royal Statistical Society, Series B* **34**, 216–217.
- Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.
- Cox, D. R. (1972). Regression models and life-table (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81**, 1879–1886.
- Greipp, P. R., Leong, T. and et al. (1998). Plasmablastic morphology-an independent prognostic factor with clinical and laboratory correlation. *Blood* **91**, 2501–2507.
- Lee, J. (1998). Calibration. *Encyclopedia of Biostatistics* **1**, 461–466.
- Lin, D. Y., Fleming, T. R. and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81.
- Oken, M. M., Leong, T., Lenhard, R. E., Greipp, P. R., Kay, N. E., Van Ness, B., Keimowitz, R. M. and Kyle, R. A. (1999). The addition of interferon or high dose cyclophosphamide to standard chemotherapy in the treatment of patients with multiple myeloma. *Cancer* **86**, 957–968.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York.
- Sun, J. (1993). Tail probabilities of the maxima of gaussian random fields. *Annals of Probability* **21**, 34–71.

- Sun, J., Loader, C. and McCormick, W. P. (2000). Confidence bands in generalized linear models. *Annals of Statistics* **28**, 429–460.
- Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine* **11**, 1871–1879.

APPENDIX

First, we show how to obtain an approximation to the distribution of

$$\sup_{\mathbf{z} \in \mathcal{H}} \frac{(1, \mathbf{z}')G}{\sqrt{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'}}, \quad (6.1)$$

where $\mathbf{z} \in \mathcal{H}$ is defined by a set of linear (equality and/or inequality) constraints of $\mathbf{z}$. For any given realization g of G , let $y = (1, \mathbf{z}')g$. Then, (6.1) becomes

$$\sup_{-\infty < y < \infty} \frac{y}{h(y)}, \quad (6.2)$$

where

if $y \geq 0$,

$$h(y) = \inf_{\mathbf{z}} \{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'\} \quad (6.3),$$

if $y < 0$,

$$h(y) = \sup_{\mathbf{z}} \{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'\} \quad (6.4).$$

Here, inf and sup are taken over $\mathbf{z} \in \mathcal{H} \cap \{(1, \mathbf{z}')g = y\}$. Note that both (6.3) and (6.4) can be obtained via the standard quadratic programming with linear constraints. Commercial software is widely available for implementing such optimization problems (see the website: www.ece.northwestern.edu/OTC). When Σ is strictly positive definite, any local minimum of $(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'$ is the global minimum. Therefore, the numerical search for $h(y)$ in (6.3) is a rather trivial task. On the other hand, locating a global maximizer for $(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'$ in (6.4) may not be that straightforward due to the fact that there are possible local

minima. Note that if for any non-negative y , the feasible set for (6.3) is not empty, that is, the constrain $\mathbf{z} \in \mathcal{H} \cap \{(1, \mathbf{z}')g = y\}$ does not give us an empty set of $\mathbf{z}$, we do not have to consider (6.4) to obtain (6.2). Empirically, we find that it is rather rare that all the feasible sets for (6.3) are empty. In any event, since $\frac{y}{h(y)}$ is a function of a single variable y , its sup can be obtained numerically without much difficulty. One can then simulate a large number of realizations g from G and obtain an approximation to the distribution of (6.1). Note that the critical value c_α may also be approximated via the techniques recently developed for obtaining tail probabilities of the maxima of Gaussian random fields (Sun (1993); Sun, Loader and McCormick (2000)).

When all the covariates are continuous, any bounded support of the covariates can be defined by a linear inequality constrain of $\mathbf{z}$. Now, suppose that the first component of $\mathbf{Z}$ is discrete with two possible values, say, 0 and 1, and the rest of covariates are continuous. Then, the corresponding $\sup_{\mathbf{z} \in \mathcal{H}} \{W(\mathbf{z})/\sigma(\mathbf{z})\}$ can be obtained by taking the maximum of two sup-statistics similar to (6.1), that is, $\sup_{\mathbf{z} \in \mathcal{H}_l} \frac{(1, \mathbf{z}')G}{\sqrt{(1, \mathbf{z}')\Sigma(1, \mathbf{z}')'}}$, where $\mathcal{H}_l = \mathcal{H} \cap \{(1, 0, \dots, 0)\mathbf{z} = l\}$ and $l = 0, 1$. For the general case with a mixture of discrete (say, a total of m possible combinations) and continuous covariates, the $\sup_{\mathbf{z} \in \mathcal{H}} W(\mathbf{z})/\sigma(\mathbf{z})$ can be obtained by taking the maximum of m sup-statistics similar to (6.1), each of them has an extra linear equality constrain defined by a set of specific values generated from all the discrete covariates.

Now, consider (6.1) with $\mathcal{H} = R^d$. Let $\mathbf{v} = \Sigma^{1/2}(1, \mathbf{z}')'$, $\mathbf{h} = \Sigma^{-1/2}g$. Then (6.1) = $\sup_{\mathbf{v}: \mathbf{v}'\mathbf{s}=1} \frac{\mathbf{v}'\mathbf{h}}{\sqrt{\mathbf{v}'\mathbf{v}}}$, where $\mathbf{s}' = (s_1, \dots, s_{d+1})$ is the first row of the matrix $\Sigma^{-1/2}$. Recall that $\mathbf{v}'\mathbf{h} = y$. It follows from 1f.1.3 of Rao (1973) (p.60) that the minimum value $h(y)$ of $\mathbf{v}'\mathbf{v}$ is

$$(1, y) \begin{pmatrix} \mathbf{s}'\mathbf{s} & \mathbf{s}'\mathbf{h} \\ \mathbf{s}'\mathbf{h} & \mathbf{h}'\mathbf{h} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ y \end{pmatrix} = \frac{\sum_{i=1}^{d+1} h_i^2 - 2(\sum_{i=1}^{d+1} s_i h_i)y + (\sum_{i=1}^{d+1} s_i^2)y^2}{\sum_{i=1}^{d+1} s_i^2 \sum_{i=1}^{d+1} h_i^2 - (\sum_{i=1}^{d+1} s_i h_i)^2},$$

where h_i is the i th component of $\mathbf{h}$. This implies that the maximum value of $y/h(y)$ is

the maximum value of

$$\frac{y\sqrt{\sum_{i=1}^{d+1} s_i^2 \sum_{i=1}^{d+1} h_i^2 - (\sum_{i=1}^{d+1} s_i h_i)^2}}{\sqrt{\sum_{i=1}^{d+1} h_i^2 - 2(\sum_{i=1}^{d+1} s_i h_i)y + (\sum_{i=1}^{d+1} s_i^2)y^2}}$$

over $y > 0$, which is

$$\begin{cases} \sqrt{\sum_{i=1}^{d+1} h_i^2 - \frac{(\sum_{i=1}^{d+1} s_i h_i)^2}{\sum_{i=1}^{d+1} s_i^2}} & \text{if } \sum_{i=1}^{d+1} s_i h_i < 0 \\ \sqrt{\sum_{i=1}^{d+1} h_i^2} & \text{if } \sum_{i=1}^{d+1} s_i h_i \geq 0 \end{cases}$$

Since $\Sigma^{1/2}G$ is the standard normal, it is easy to show that the above quantity or (6.2) is a realization of a mixture of χ_d and χ_{d+1} with a mixing probability of 0.5. Note that one may use the $100(1 - \alpha)$ th percentile of this mixing distribution to be an upper bound of the cutoff point c_α for the case with any type of covariate vector $\mathbf{Z}$.

Figure 1. The graphical display for point and interval estimates for Ω_p

Figure 2. Kaplan-Meier estimates of survival probabilities for three treatment groups in the myeloma study

$$(a) \Omega_p^* = \{\mathbf{z} : \text{pr}(T \leq 42 | \mathbf{z}) \geq 0.7\}$$

Figure 3 Estimated regions of predictors for short-term survivors with multiple myeloma.

(b). $\Theta_{\tau}^* = \{\mathbf{z} : \text{the median of } T \leq 36\}$

(a). $\Omega_p = \{\mathbf{z} : \text{pr}(T \geq 42|\mathbf{z}) \geq 0.7\}$

Figure 4. Estimated regions of predictors for long-term survivors with the multiple myeloma.

(b). $\Theta_\tau = \{\mathbf{z} : \text{the median of } T \geq 48\}$

Table 1

The empirical coverage levels of interval estimators for Ω_p

n	no censoring		20% censoring	
	0.95	0.90	0.95	0.90
100	0.91	0.84	0.91	0.84
250	0.92	0.86	0.92	0.86
450	0.94	0.88	0.94	0.88
650	0.94	0.89	0.94	0.89
850	0.94	0.89	0.94	0.89
1050	0.94	0.89	0.94	0.89
1250	0.95	0.89	0.94	0.89