Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 38

Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models

Hajime Uno* Tianxi Cai†
Lu Tian‡ L.J. Wei**

*Kitasato University, unoh@pharm.kitasato-u.ac.jp
†Harvard University, tcai@hsph.harvard.edu
‡Northwestern University, lutian@northwestern.edu
**Harvard University, wei@sdac.harvard.edu

Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models

Hajime Uno, Tianxi Cai, Lu Tian, and L.J. Wei

Abstract

Suppose that we are interested in establishing simple, but reliable rules for predicting future t-year survivors via censored regression models. In this article, we present inference procedures for evaluating such binary classification rules based on various prediction precision measures quantified by the overall misclassification rate, sensitivity and specificity, and positive and negative predictive values. Specifically, under various working models we derive consistent estimators for the above measures via substitution and cross validation estimation procedures. Furthermore, we provide large sample approximations to the distributions of these nonsmooth estimators without assuming that the working model is correctly specified. Confidence intervals, for example, for the difference of the precision measures between two competing rules can then be constructed. All the proposals are illustrated with two real examples and their finite sample properties are evaluated via a simulation study.

Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models

By Hajime Uno
Division of Biostatistics, School of Pharmaceutical Sciences, Kitasato University, Tokyo, Japan, 108-8641
unoh@pharm.kitasato-u.ac.jp
Tianxi Cai
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115
tcai@hsph.harvard.edu
Lu Tian
Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611
lutian@northwestern.edu
and L. J. Wei
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115
wei@sdac.harvard.edu

Abstract

Suppose that we are interested in establishing simple, but reliable rules for predicting future t-year survivors via *censored* regression models. In this article, we present inference procedures for evaluating such binary classification rules based on various prediction precision measures quantified by the overall misclassification rate, sensitivity and specificity, and positive and negative predictive values. Specifically, under various *working* models we derive consistent estimators for the above measures via substitution and cross validation estimation procedures. Furthermore, we provide large sample approximations to the distributions of these nonsmooth estimators without assuming that the working model is correctly specified. Confidence intervals, for example, for the difference of the precision measures between two competing rules can then be constructed. All the proposals are illustrated with two real examples and their finite sample properties are evaluated via a simulation study.

Key words: Cross validation; Gene expression; Model selection; Positive and negative predictive values; Prediction error; ROC curve; Survival analysis.

1. INTRODUCTION

Suppose that we are interested in establishing reliable and parsimonious classification rules for predicting future patients' survival based on the data collected from a current study. Typically the data consist of a set of survival times, possibly censored, and their corresponding "baseline" covariates. To predict covariate specific survival, we fit the data with a parametric or semi-parametric regression model, for example, the proportional hazards model (Cox 1972), the accelerated failure time model (Wei 1992; Kalbfleisch and Prentice 2002, chap. 7), or the transformation model (Cheng, Wei and Ying 1995). With this fitted model, one can estimate the survival function for a future subject using its covariate information and then predict, for example, whether the patient would survive more than t years. Oftentimes the aforementioned survival models assume that the covariate effects on the patient's survival or hazard function are constant over the entire follow-up study period. This modeling assumption may be reasonable for a global assessment of the covariate effects on survival. From the prediction point of view, however, a good classification rule for predicting short term survivors may perform poorly for predicting long term survivors. We will address this issue in this article via a rather simple time varying binary regression modeling approach.

When there is no censoring, standard methods for binary outcomes such as logistic regression, CART, neural networks and discriminant analysis (Breiman, Friedman, Olshen and Stone 1984; McLachlan 1992; Ripley 1996) may be used to construct prediction rules. To evaluate a classifier, various prediction precision measures, which quantify the discordance or concordance between the observed and the predicted outcomes, have been utilized, for example, the probability scores (Brier 1950; Spiegelhalter 1986), the explained variation (Korn and Simon 1990; Mittlbock and Schemper 1996), the overall misclassification rate (OMR),

the sensitivity (SE) and specificity (SP), and positive and negative predictive values (PPV & NPV) (Zhou, Obuchowski and McClish 2002; Pepe 2003).

In the presence of censoring, especially when the censoring support is shorter than its survival counterpart, very few methods are available for constructing and evaluating t-year survivor prediction rules. For the case of a univariate covariate, Heagerty, Lumley and Pepe (2000) proposed non-parametric estimators for the SE and SP, and Moskowitz and Pepe (2004) considered marginal regression models for comparing the PPV and NPV of two competing prediction rules based on hypothesis testing. When there are multiple covariates involved, Heagerty and Zheng (2005) developed a prediction rule through a proportional hazards model with time varying coefficients. Recently Zheng, Cai and Feng (2006) proposed a prediction rule based on a time varying logistic regression model and evaluated its overall accuracy through simulation. Note that all the aforementioned procedures are derived under the assumption that the working model is correctly specified. Moreover, there are no theoretically justified methods for constructing interval estimates of the prediction precision measures when more than one covariate is available.

In this article, we propose classification rules for predicting t-year survival based on a class of simple *working* models which only relate the covariates to the patient's t-year survival probability. Under the assumption that the censoring distribution of the current study is independent of the covariates or can be modeled reasonably well, for each prediction rule we show how to consistently estimate its OMR, SE, SP, PPV and NPV. Note that most existing estimation procedures for the commonly used survival models may not be able to provide such consistent estimators when the model is incorrectly specified (O'Quigley and Xu 2001). In addition to providing point estimates for the prediction precision measures, we also derive the

large sample distribution of the proposed estimators. Furthermore, since these estimators are not smooth, a perturbation-resampling technique is utilized to approximate their distributions without involving any nonparametric density-like function estimates. Base on these large sample approximations, confidence intervals for the OMR, SE, SP, PPV and NPV, or functions thereof, can be constructed accordingly, which provide much more information than their point estimate counterparts for evaluating regression models and their resulting prediction rules.

If the same dataset is used to construct the prediction rules and evaluate their performance, the above substitution or "apparent error" estimates may be biased (Efron 1983, 1986) especially when the sample size is not large with respect to the number of the covariates in the model. To reduce the potential bias of the apparent error, methods such as cross-validation, bootstrap, and covariance penalties have been proposed for certain regression models with non-censored data (Mallows 1973; Efron 1986; Shao 1996; Efron and Tibshirani 1997; Ye 1998; Tibshirani and Knight 1999; Efron 2004). In this article, we also study properties of bias corrected estimators for the OMR, SE, SP, PPV and NPV via various cross validation schemes. Lastly, we provide interval estimates for the difference of the prediction precision measures between two competing classification rules or models. Note that our procedures can be easily generalized to the case when we are interested in making joint inferences about the performance of prediction rules for a set of time points t. All the proposals are illustrated and evaluated via two examples and a simulation study.

2. EVALUATING PREDICTION RULES FOR t-YEAR SURVIVORS BASED ON OVERALL MISCLASSIFICATION RATE

Let T be a continuous failure time and $\tilde{Z}$ be a set of bounded potential predictors. Also,

let C be the corresponding censoring variable. Assume that T and C are independent and the survival function $G(\cdot)$ of C is free of $\tilde{Z}$. Let $\{(T_i, \tilde{Z}_i, C_i), i = 1, \cdots, n\}$ be n independent copies of $(T, \tilde{Z}, C)$. For the ith subject, we only observe $(X_i, \tilde{Z}_i, \Delta_i)$, where $X_i = \min(T_i, C_i)$, $\Delta_i = I(X_i = T_i)$, and $I(\cdot)$ is the indicator function.

Suppose that based on the data $\{(X_i, \tilde{Z}_i, \Delta_i), i = 1, \cdots, n\}$, we are interested in establishing a rule which can "accurately" predict whether the survival time T^0 of a future subject with $\tilde{Z} = \tilde{Z}^0$ is shorter than t-year or not, where t is a pre-specified time point and $\text{pr}(X > t) > 0$. To this end, let Z, a function of $\tilde{Z}$, be a p-dimensional vector with the first component being one and consider the following *working* model

$$\text{pr}(T < t|Z) = g(\beta' Z), \tag{2.1}$$

where $g(\cdot)$ is a known strictly increasing, differentiable function and β is a p-dimensional vector of unknown parameters. Note that if we assume Model (2.1) for all $t \geq 0$ and the *first* component of β depends on t, (2.1) is called the linear transformation model (Cheng et al. 1995). In particular, if $g(\cdot)$ is $1 - \exp(-\exp(\cdot))$, (2.1) is the proportional hazards model. On the other hand, if $g(\cdot)$ is the anti-logit function, (2.1) is the so-called proportional odds model. In this article, for each time point t of interest, we let $g(\cdot)$ and all the components of β in (2.1) depend on t. With this more flexible modeling, we may find, for example, that a good prediction rule for short term survivors may be quite different from that for long term survivors.

Suppose that β in (2.1) is estimated by $\hat{\beta}$ based on the data $\{(X_i, Z_i, \Delta_i)\}$. For a future subject with a covariate vector $Z = Z^0$, consider a class of binary prediction rules indexed by c: $I(g(\hat{\beta}' Z^0) \geq c)$, where $0 < c < 1$. For example, if $c = .5$, and $g(\hat{\beta}' Z^0) \geq .5$, we predict that

this subject would die by time t. To evaluate this class of prediction rules, consider the OMR

$$D_n(c) = E|I(T^0 < t) - I(g(\hat{\beta}' Z^0) \geq c)|,$$

where the expectation is taken over $\{(X_i, Z_i, \Delta_i)\}$ and (T^0, Z^0). Suppose that as $n \to \infty$, $\hat{\beta}$ converges to a constant vector β_0, which is free of $G(\cdot)$, and $D_n(c)$ goes to

$$D(c) = E|I(T^0 < t) - I(g(\beta_0' Z^0) \geq c)|. \tag{2.2}$$

Now, let c_0 be a minimizer of $D(c)$ for $0 \leq c \leq 1$, and let $D(c_0) = D_0$, which does not depend on the nuisance censoring distribution. To evaluate the adequacy of Model (2.1) as a prediction tool, we need to estimate D_0 and c_0.

When a working survival model is not correctly specified, it is not clear that the existing estimator of the vector of regression parameters would converge to a constant vector, as $n \to \infty$. Moreover, even when the estimator is stabilized for large n, its limit may depend on the distribution of the nuisance censoring variable C. Consequently, the corresponding $D_n(c)$ converges to a quantity, which may also depend on the censoring and may not be a meaningful criterion for evaluating prediction rules. Here, we propose a simple estimator $\hat{\beta}$ for β in the working model (2.1), which converges to a constant vector β_0 that is free of the censoring distribution.

Our estimator $\hat{\beta}$ is based on the following estimating function (Zheng et al. 2006)

$$U(\beta) = n^{-1} \sum_{i=1}^{n} \frac{w_i}{\hat{G}(X_i \wedge t)} Z_i \{I(X_i < t) - g(\beta' Z_i)\}, \tag{2.3}$$

where $w_i = I\left(T_i \wedge t \leq C_i\right) = I\left(X_i \leq t\right) \Delta_i + I\left(X_i > t\right)$, and $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$. Note that if the ith subject is censored before time t, $w_i = 0$. On the other hand, such censored observations are included in the construction of $\hat{G}(\cdot)$. Note that conditional on

(T_i, Z_i), the expected value of $w_i\{G(X_i \wedge t)\}^{-1}$ is one. Therefore, conditional on $\{(T_i, Z_i)\}$, asymptotically the expected value of $U(\beta)$ is $u(\beta) = E[Z\{I(T < t) - g(\beta' Z)\}]$, which is free of the censoring variable C.

Under a rather mild condition that there does not exist a β such that $P(\beta' Z_1 > \beta' Z_2 \mid T_1 < t \leq T_2) = 1$, using a similar argument given in Appendix A of Tian, Cai, Goetghebeur and Wei (2005), one can show that $u(\beta) = 0$ has a unique solution, say, β_0. Moreover, if there does not exist a β such that $\Delta_i I(X_i < t \leq X_j) = 1$ implies that $I(\beta' Z_i \leq \beta' Z_j) = 1$, for any pair $1 \leq i \leq j \leq n$, $U(\beta) = 0$ has a unique solution $\hat{\beta}$ for any finite n. Since $\hat{G}(s)$ converges uniformly to $G(s)$, for $s \leq t$, it follows from the uniform law of large numbers (Pollard 1990, pp. 41) that $U(\beta)$ is uniformly convergent to $u(\beta)$ in probability around the neighborhood of β_0. This implies that $\hat{\beta}$ converges to β_0, in probability, as $n \to \infty$ even when model (2.1) is not correctly specified.

Now, to estimate $D(c)$, first consider the so-called apparent error (Davison and Hinkley 1997, pp. 292)

$$\hat{D}(c) = \frac{1}{n} \sum_{i=1}^{n} \frac{w_i}{\hat{G}(X_i \wedge t)} |I(X_i < t) - I(g(\hat{\beta}' Z_i) \geq c)|. \tag{2.4}$$

Let $\hat{c}$ be a minimizer of $\hat{D}(c)$, for $0 \leq c \leq 1$. In Appendix A, under the mild condition that $\text{pr}(T < t |\ \beta_0' Z = y)$ is strictly increasing in y in the support of $\beta_0' Z$, we show that $\hat{c}$ and $\hat{D}(\hat{c})$ are consistent with respect to c_0 and D_0. Note that one can check this condition empirically by estimating $\text{pr}(T < t |\ \beta_0' Z = y)$ via a nonparametric function estimate based on the "data" $\{(X_i, \Delta_i, \hat{\beta}' Z_i), i = 1, \cdots, n\}$.

To make further inferences about D_0, consider a standardized transformation of $\hat{D}(\hat{c})$:

$$n^{\frac{1}{2}}\{\text{log(-log)}(\hat{D}(\hat{c})) - \text{log(-log)}(D_0)\}, \tag{2.5}$$

which is asymptotically equivalent to

$$\{\hat{D}(\hat{c})\log(\hat{D}(\hat{c}))\}^{-1}W,$$

where

$$W = n^{\frac{1}{2}}\{\hat{D}(\hat{c}) - D_0\}.$$

In Appendix B, we show that W is asymptotically equivalent to $n^{\frac{1}{2}}\{\hat{D}(c_0) - D_0\}$ and converges in distribution to a normal with mean 0. However, the variance of W, which involves unknown density-like functions, is difficult to estimate well directly. One may use a perturbation-resamling method to obtain a good approximation to the distribution of W. To be specific, let $\{V_i, i = 1, \ldots n\}$ be n independent copies of a random variable V from a known distribution with mean one and variance one. Let $D^*(c)$ be a perturbed version of $\hat{D}(c)$, where

$$D^*(c) = n^{-1}\sum_{i=1}^{n}\frac{w_i}{G^*(X_i \wedge t)}|I(X_i < t) - I(g(Z_i'\beta^*) \geq c)|V_i, \tag{2.6}$$

and $G^*(\cdot)$ and β^* are the corresponding perturbed versions of $\hat{G}(\cdot)$ and $\hat{\beta}$. To construct $G^*(\cdot)$, we use the martingale representation formula for the Kaplan-Meier estimate (Fleming and Harrington 1991, pp. 98). Specifically, for $t > 0$, the unconditional distribution of $\hat{G}(t) - G(t)$ can be approximated by the conditional distribution (given the data) of

$$-\hat{G}(t)\sum_{i=1}^{n}V_i\int_0^t\{\sum_{j=1}^{n}I(X_j \geq s)\}^{-1}d\hat{M}_i(s),$$

where $\hat{M}_i(t) = I(X_i \leq t, \Delta_i = 0) - \int_0^t I(X_i \geq s)d\hat{\Lambda}(s)$, and $\hat{\Lambda}(\cdot)$ is the standard Nelson-Aalan estimator of the cumulative hazard function for the censoring variable C. It follows that

$$G^*(t) = \hat{G}(t) - \hat{G}(t)\sum_{i=1}^{n}V_i\int_0^t\{\sum_{j=1}^{n}I(X_j \geq s)\}^{-1}d\hat{M}_i(s). \tag{2.7}$$

To obtain a perturbed β^*, we solve the equation

$$U^*(\beta) = n^{-1} \sum_{i=1}^{n} \frac{w_i}{G^*(X_i \wedge t)} Z_i \{I(X_i < t) - g(\beta' Z_i)\} V_i = 0. \tag{2.8}$$

Note that since $U(\beta)$ is a differentiable function in β, an alternative way to obtain $(\beta^* - \hat{\beta})$ is by perturbing the first order expansion of $n^{1/2}(\hat{\beta} - \beta_0)$. It follows from similar arguments given in Park and Wei (2003) or Cai, Tian and Wei (2005) that the distribution of (2.5) can be approximated by the conditional distribution of

$$\{\hat{D}(\hat{c}) \log(\hat{D}(\hat{c}))\}^{-1} W^*,$$

given the data, where $W^* = n^{1/2}\{D^*(\hat{c}) - \hat{D}(\hat{c})\}$.

In practice, one may generate a large number M of random samples W^* to approximate the distribution of W. Confidence interval estimates of D_0 can then be constructed accordingly via (2.5). Note that if we let $\{V_i, i = 1, \ldots, n\}$ be the multinomial random vector with size n and cell probability of n^{-1}, the above resampling method is similar to the standard Efron's bootstrapping (Efron 1982). However, it is not clear how to justify the large sample approximation to the distribution of W using perturbation with such dependent V's.

When the sample size n is not large with respect to the dimension of the covariate vector Z, one may use cross-validation methods to estimate the prediction error D_0. To this end, we first consider the commonly used K-fold cross-validation, which randomly splits the data into K disjoint sets of about equal size and label them as $\mathcal{I}_k, k = 1, \cdots, K$. For each k, an estimate $\hat{\beta}_{(-k)}$ for β via (2.3) is obtained based on all observations which are not in $\mathcal{I}_k$. We then compute the predicted error estimate $\hat{D}_{(k)}(c)$ via (2.4) based on observations in $\mathcal{I}_k$. Then, an average prediction error estimate for $D(c)$ is

$$\hat{\mathcal{D}}(c) = K^{-1} \sum_{k=1}^{K} \hat{D}_{(k)}(c). \tag{2.9}$$

Let $\hat{c}_v$ be a minimizer of $\hat{\mathcal{D}}(c)$, for $0 \leq c \leq 1$. When K is small with respect to n, it is straightforward to show that $\hat{c}_v$ and $\hat{\mathcal{D}}(\hat{c}_v)$ are consistent for c_0 and D_0, respectively. Moreover, in Appendix C, we show that the standardized $\hat{\mathcal{D}}(\hat{c}_v)$

$$\mathcal{W} = n^{1/2}\{\hat{\mathcal{D}}(\hat{c}_v) - D_0\}$$

has the same limiting distribution as that of W based on the apparent error. Therefore, one may use the standard error estimate of the apparent error to construct interval estimates for D_0, which are centered around the cross validation estimate.

For a general cross-validation, let n_t and n_v be the sizes of the training and validation sub-samples, where n/n_v is roughly a fixed positive integer, and n_t and $n_v \to \infty$, as $n \to \infty$. We randomly choose a training set to obtain an estimate for β via (2.3), then compute $\hat{D}(c)$ in (2.4) with the validation set. We repeat this process by taking a fresh random training and validation partition. Let $\hat{\mathbb{D}}(c)$ be the average of all $\hat{D}(c)$ over the entire set of possible random splits of the training-validation sub-samples. Let $\hat{c}_{rv}$ be a minimizer of $\hat{\mathbb{D}}(c)$. In Appendix C, we show that the distribution of $n^{1/2}(\hat{\mathbb{D}}(\hat{c}_{rv}) - D_0)$ is the same as that of W in the limit, and thus can be approximated well by that of W^*.

Now, suppose that we are interested in comparing two working models (2.1) with possibly different covariate vectors, say, $Z_{(l)}, l = 1, 2$. To this end, all the above notations are sub-indexed by $l, l = 1, 2$. For example, for Model l with the optimal cut-off point $c_0 = c_{l0}$, the link function in Model (2.1) is $g_l(\cdot)$. Let $\tau = D_2(c_{20}) - D_1(c_{10})$ and $\hat{\tau} = \hat{D}_2(\hat{c}_2) - \hat{D}_1(\hat{c}_1)$. Then, the distribution of $W_\tau = n^{1/2}(\hat{\tau} - \tau)$ is approximately normal with mean 0. Now, let $\tau^* = D_2^*(\hat{c}_2) - D_1^*(\hat{c}_1)$. Note that for $D_1^*(\cdot)$ and $D_2^*(\cdot)$, we need to use the same set of perturbation variables $\{V_i, i = 1, \cdots\}$ in (2.6), (2.7) and (2.8). Then, the distribution of W_τ

can be approximated well by the conditional distribution of $W_\tau^* = n^{1/2}(\tau^* - \hat{\tau})$. Confidence intervals for τ can then be constructed via this approximation. For the aforementioned K-fold and random cross-validation schemes, we can construct the corresponding estimates $\hat{\mathcal{D}}_l(\hat{c}_v)$ and $\hat{\mathbb{D}}_l(\hat{c}_{rv}), l = 1, 2$ to make inferences about τ.

Now, we use two examples to illustrate our proposals. The first one is from the well-known Mayo primary biliary cirrhosis study (Fleming and Harrington 1991, app. D). The dataset utilized here consists of 418 patient records, each of which contains the survival time and seventeen potential prognostic factors. To simplify the illustration, we considered only five covariates: age, log(albumin), log(bilirubin), edema and log(protime), which were selected as the most important predictors based on a Cox regression model (Dickson, Fleming, Grambsch, Fisher and Langworthy 1989; Fleming and Harrington 1991, pp. 195). Suppose that we are interested in establishing prediction rules for ten-year survivors based on the above five covariates. First we considered two different models (2.1) with $g(y) = 1 - \exp\{-\exp(y)\}$ to fit the data. The first model uses age only, and the second one takes the above five covariates additively. With apparent errors $\hat{D}(c)$, for all cases studied here, $\hat{c} \approx .5$. We report $\hat{D}(\hat{c})$ and the corresponding standard error estimates for D_0 in Table 1. All standard error estimates were constructed based on $M = 2000$ sets of $\{V_i\}$, where V is the unit exponential. In the Table, we also report the point estimates based on the 10-fold and random cross validation procedures. For the random cross validation, we let the training set size be $2n/3$ for each of 200 iterations. For Model II, the apparent error estimate appears noticeably small compared with its random cross validation counterpart. From the Table, based on the random cross validation point estimates, 95% intervals for the misclassification rate for Model I and II are (.24, .44) and (.14, .31), respectively. We also report 95% confidence intervals for the difference

of error rates between two fitted models in Table 1. For example, the interval estimate for the difference of two rates D_0, Model I minus II, is (.03, .21), indicating that Model II, which includes clinical biomarkers, is better than Model I with respect to the 10-year survival prediction. On the other hand, the degree of improvement ranges from 3% to 21%, reflecting rather large sampling variation.

It is interesting to note that for Model II, unlike the results from the standard Cox model fitting, edema and log(protime) are not statistically significant (the p-values for testing no covariate effect are .37 and .23, respectively). To explore if these two clinical markers are needed for prediction, we fit the data with Model III, which consists of three covariates: age, log(bilirubin) and log(albumin). The resulting points and the standard error estimates are reported in Table 1. The 95% interval estimate for the difference of the error rates between Models III and II is (-.03, .05), indicating that edema and protime have no added value over the other three covariates for predicting ten-year survivors with respect to the overall prediction rate.

Now, it is also interesting to investigate that a statistically significant covariate may not add any substantial value for prediction. To this end, we create Model IV by deleting a highly, statistically significant covariate, log(albumin), from Model III. We report the point and standard error estimates for this model in Table 1. The 95% confidence interval for the difference of the OMR between Model IV and III is $(-.07, .06)$, indicating that age and bilirubin appears to be sufficient for predicting the 10-year survivors with respect to the OMR.

The second example is from a recent study on prognostication in breast cancer with microarray gene expression data (van de Vijver et al. 2002). There are 295 breast cancer patients in the study. For each patient, we have her survival time, baseline lymph-node status, estro-

gen receptor status, and "gene signature score" *(http://www.rii.com/publications/).* The gene score, which is continuous and between 0 and 1, was derived from the gene expression data based on 70 selected genes via a supervised classification algorithm for predicting the distant metastases within five years after patient's surgery. A patient with a high score is expected to have a long survival time. One of the clinical objectives for this study is to identify future patients via gene score values, who may benefit from potentially toxic, adjuvant systemic therapy. van de Vijver et al. (2002) proposed a binary prediction rule based solely on the gene score. For illustration, suppose that we are interested in predicting ten-year survivors and consider three models (2.1). The first one does not use any covariate, the second model uses the above two clinical marker values, and the third one uses clinical markers and also the gene score. Again, for all cases studied here, $\hat{c} \approx .5$. Since the ten-year survival rate for this study is approximately .7, Model I essentially produces a rule which predicts all future patients would survive beyond ten years. The error rate for this naive rule is .3. In Table 2, we present the apparent errors $\hat{D}(\hat{c})$ and the corresponding standard error estimates for D_0. For the present case, the estimates based on cross validation are quite similar to the apparent errors. In the Table, we also include the estimated error rate for the prediction rule proposed by van de Vijver et al. (2002) for comparisons. With respect to the overall misclassification rate, it is interesting to note that the prediction rules, which utilize the "baseline" clinical or gene expression information, do not perform better than the aforementioned naive rule (Model I). In fact, the error rate for the rule by van de Vijver et al. (2002) is 35%, which is higher than the ten-year mortality rate of 30%. Since for the present case, it is critical to accurately identify future breast cancer patients who would likely die before ten years after surgery, the overall misclassification rate may not be a good criterion for evaluating prediction rules. We

discuss other evaluation criteria in the next two sections. Note that for both examples, for all models considered here the nonparametric function estimates for $\text{pr}(T < 10|\ \beta_0' Z)$ appear to be monotone in $\beta_0' Z$.

To examine finite sample properties of the proposed estimation procedures based on $\hat{D}(\hat{c})$, $\hat{\mathcal{D}}(\hat{c}_v)$ and $\hat{\mathbb{D}}(\hat{c}_{rv})$, we conducted a simulation study under a practical setting. Specifically, we mimicked the Mayo study to generate realizations of $T, \tilde{Z}$ and C. Here, $\tilde{Z} = (1, Z_{cov}')'$ and Z_{cov} is a multivariate normal whose mean and covariance matrix are estimated based on the 282 completely observed vectors consisting of age, log(bilirubin), log(albumin), log(sgot), log(protime), log(cholesterol), and log(copper) from the Mayo study. Now, let $Z_0' =$ (age, log(bilirubin), log(albumin)). For each realized Z_0 from the above normal, the survival time T is generated via an exponential with a scale parameter of $\exp(b' Z_0)$, where b is estimated by the above 282 observed censored failure times and their corresponding observed covariate vectors of age, log(bilirubin) and log(albumin) with this exponential model. Lastly, the censoring distribution of C is the Kaplan-Meier estimate from the Mayo study. Note that the proportion of the above Mayo patients among complete cases whose survival times were censored at year 10 is about 52%.

In our numerical study, we considered six working models (2.1). For Model I, we let $g(\cdot)$ be the inverse function of 1-log(-log)$(\cdot)$ and $Z = Z_0$. Note that Model I is the correct model for $\text{pr}(T < t|\ Z_0)$. For Model II, we deleted the covariate log(albumin) from the above Z. For Model III, we let $Z =$ (1, age, bilirubin, albumin), a case with wrong transformations of covariates. For Model IV, we considered the case with a wrong link function, that is, we let $1 - g(\cdot)$ be the anti-logit function with $Z = Z_0$. In Model V, we let g be the above wrong link, and also ignored the log-transformation of bilirubin and albumin. For Model VI, we

considered an "overfitting" case, that is, we let $Z = \tilde{Z}$ with the correct link function.

For each working model, we generated 100,000 realizations of (T, Z) to obtain its model-specific β_0 and used another fresh 100,000 realizations to estimate the true OMR D_0. Then, we generated 2000 sets of realizations $\{T_i, C_i, \tilde{Z}_i, i = 1, \cdots, n\}$ from the aforementioned true model and obtained 2000 sets of realized $\hat{D}(\hat{c})$, $\hat{\mathcal{D}}(\hat{c}_v)$ and $\hat{\mathbb{D}}(\hat{c}_{rv})$ with $n_t = 2n/3$. Note that for the random cross validation, we used 100 random splits of the sample. Based on these realized point estimates, we obtained the average bias and root mean-square error (RMSE). Furthermore, we obtained 2000 standard error estimates. Each of these estimates is based on 2000 perturbed W^*. Then, for each of the above three types of point estimates, we constructed 2000 95% confidence intervals for D_0. In Table 3, we reported the results with $n = 300$ and $t = 10$ years under the heading "Observed censoring". The cross-validation indeed reduces bias of the apparent error. However, the bias of the apparent error seems rather small with respect to the true error D_0 for each working model. Moreover, with respect to RMSE, the three estimation procedures are compatible with each other. On the other hand, the empirical coverage level of the confidence interval centered about the apparent error tends to be lower than its nominal counterpart. In Table 3, we also report results for the case that there is no censoring involved. Again, with respect to the coverage probability, the interval estimate centered about the cross-validation point estimate appears to be better than its apparent error counterpart.

3. EVALUATION BASED ON SENSITIVITY AND SPECIFICITY

To evaluate a prediction rule for a binary outcome, one may also consider its sensitivity and specificity. For the prediction rule: $I(g(\hat{\beta}'Z) > c)$, the sensitivity is SE(c) =pr$(g(\beta_0'Z^0) \geq$

$c|\ T^0 < t)$ and the specificity is SP(c)= $\text{pr}(g(\beta_0' Z^0) < c|\ T^0 \geq t)$. These conditional probabilities can be estimated consistently by

$$\hat{\text{SE}}(c) = \frac{\sum_{i=1}^n w_i \{\hat{G}(t \wedge X_i)\}^{-1} I(g(\hat{\beta}' Z_i) \geq c, X_i < t)}{\sum_{i=1}^n w_i \{\hat{G}(t \wedge X_i)\}^{-1} I(X_i < t)}, \tag{3.1}$$

and

$$\hat{\text{SP}}(c) = \frac{\sum_{i=1}^n w_i \{\hat{G}(t \wedge X_i)\}^{-1} I(g(\hat{\beta}' Z_i) < c, X_i \geq t)}{\sum_{i=1}^n w_i \{\hat{G}(t \wedge X_i)\}^{-1} I(X_i \geq t)}, \tag{3.2}$$

respectively. To evaluate a specific working model (2.1), one may construct the commonly used receiver operating characteristic (ROC) curve using (3.1) and (3.2) (Heagerty and Zheng 2005). Furthermore, one can obtain the K-fold and random cross-validation estimates for SE(c) and SP(c).

To illustrate our proposal, we fitted the breast cancer data (van de Vijver et al. 2002) with Models I, II and III presented in Table 2 and then constructed the corresponding ROC curves based on $\{(1 - \hat{SP}(c), \hat{SE}(c)), 0 \leq c \leq 1\}$. These curves are presented in Figure 1. For Model I, the ROC curve only assumes a single point (the black circle). The estimated sensitivity of this naive rule is zero, which generally is not acceptable. Due to the discrete nature of the clinical marker values (both are binary), the curve for Model II only assumes three distinct values (denoted by open circles). For the prediction rule proposed by van de Vijver et al. (2002), the curve assumes only one value which is denoted by "$\ast$". Based on the ROC curves, Model III appears to be better than Models I and II. Moreover, Model III can produce a rule which has almost identical SE and SP to those proposed by van de Vijver et al. (2002). For the present case, the 10-fold and random cross validation estimates for SE(c) and SP(c) are similar to the apparent error counterparts.

To make further evaluation of a working model, for a patient with covariate Z, one may

choose the binary prediction rule: $I(g(\hat{\beta}'Z) > c^\dagger)$, such that $\mathrm{SE}(c^\dagger) = \gamma$, where $0 < \gamma < 1$ is a predetermined acceptable level for sensitivity. It is straightforward to show that when $\mathrm{pr}(T < t|\ \beta_0'Z = y)$ is positive for y in the support of $\beta_0'Z$, $c^\dagger$ is unique between 0 and 1. Let $\hat{c}^\dagger$ be a solution to $\hat{\mathrm{SE}}(c) = \gamma$. Then, $\hat{c}^\dagger$ is consistent to $c^\dagger$. Moreover, $\hat{\mathrm{SP}}(\hat{c}^\dagger)$ converges to $\mathrm{SP}(c^\dagger)$. To obtain confidence intervals for $\mathrm{SP}(c^\dagger)$, we utilize the perturbation-resampling method discussed in Section 2 to obtain an estimated standard error of $\hat{\mathrm{SP}}(\hat{c}^\dagger)$ or a transformation thereof. To be specific, the perturbed $\hat{\mathrm{SE}}(c)$ is

$$\mathrm{SE}^*(c) = \frac{\sum_{i=1}^n w_i\{G^*(t \wedge X_i)\}^{-1} I(g(Z_i'\beta^*) \geq c, X_i < t)V_i}{\sum_{i=1}^n w_i\{G^*(t \wedge X_i)\}^{-1} I(X_i < t)V_i}.$$

The perturbed $\mathrm{SP}^*(c)$ can be obtained similarly. Now, let c^* be a solution to the equation $\mathrm{SE}^*(c^*) = \gamma$. It follows from similar arguments as given for the OMR that when n is large, the distribution of $n^{1/2}(\hat{\mathrm{SP}}(\hat{c}^\dagger) - \mathrm{SP}(c^\dagger))$ can be approximated well by the conditional distribution of $n^{1/2}(\mathrm{SP}^*(c^*) - \hat{\mathrm{SP}}(\hat{c}^\dagger))$ given the data. Confidence intervals for $\mathrm{SP}(c^\dagger)$ can then be obtained via this large sample approximation. Note that for the cross validation methods discussed in Section 2, the corresponding standardized $\hat{\mathrm{SP}}(\hat{c}^\dagger)$ has the same limiting distribution as that of the above standardized apparent error. Moreover, any reasonable summary prediction precision constructed from $\mathrm{SE}(c)$ and $\mathrm{SP}(c)$, for example, the area under the ROC curve, can be estimated consistently via $\hat{\mathrm{SE}}$ and $\hat{\mathrm{SP}}$ and a large sample approximation to the resulting estimator can be obtained based on $\mathrm{SE}^*(c)$ and $\mathrm{SP}^*(c)$.

Now, we use the breast cancer data to illustrate the above procedure. Specifically, we compare Models II and III presented in Table 2. From the ROC curve for Model II in Figure 1, we let $\gamma = .69$, an attainable value for this working model empirically. The corresponding $\hat{c} = .23$ and $\hat{\mathrm{SP}}(\hat{c}) = .45$. On the other hand, for Model III with the same γ, $\hat{c} = .29$ and

$\hat{\text{SP}}(\hat{c}) = .75$. Furthermore, the 95% confidence interval for the difference of two $\text{SP}(c^\dagger)$'s (Model III minus Model II) is $(.11, .45)$, indicating that the gene score adds substantial value for predicting ten-year survivors on the top of the two clinical markers.

4. EVALUATION BASED ON POSITIVE AND NEGATIVE PREDICTIVE VALUES

For the prediction rule: $I(g(\hat{\beta}'Z) > c)$, the estimated sensitivity and specificity may be difficult to interpret in clinical practice. An alternative way is to use the positive and negative predictive values, denoted by PPV and NPV, respectively, where

$$\text{PPV}(c) = \text{pr}(T^0 < t | g(\beta_0' Z^0) \geq c)$$

and

$$\text{NPV}(c) = \text{pr}(T^0 \geq t | g(\beta_0' Z^0) < c).$$

These conditional probabilities can be consistently estimated by

$$\hat{\text{PPV}}(c) = \frac{\sum_{i=1}^n w_i \{\hat{G}(t \wedge X_i)\}^{-1} I(g(\hat{\beta}' Z_i) \geq c, X_i < t)}{\sum_{i=1}^n I(g(\hat{\beta}' Z_i) \geq c))}, \tag{4.1}$$

and

$$\hat{\text{NPV}}(c) = \frac{\sum_{i=1}^n w_i \{\hat{G}(t \wedge X_i)\}^{-1} I(g(\hat{\beta}' Z_i) < c, X_i \geq t)}{\sum_{i=1}^n I(g(\hat{\beta}' Z_i) < c))}, \tag{4.2}$$

respectively.

Note that for c close to the two ends of the interval $[0, 1]$, $\hat{\text{PPV}}(c)$ and $\hat{\text{NPV}}(c)$ may not be able to estimate their theoretical counterparts well. For each working model, one may plot the curve $\{(1 - \hat{\text{PPV}}(c), \hat{\text{NPV}}(c)), 0 < c_L \leq c \leq c_U < 1\}$, where c_L and c_U are given constants. Figure 2 gives such curves with the breast cancer gene-expression data based on Models II

and III and the prediction rule by van de Vijver et al. (2002) presented in Table 2. Here, we let $c_L = .1$ and $c_U = .8$. Note that the curve for Model II only assumes three points and its largest NPV is only .77. On the other hand, Model III appears to be more flexible and can reach rather high NPV levels. Moreover, Model III can produce a rule which matches the PPV and NPV of the scheme proposed by van de Vijver et al. (2002).

To make further inferences about evaluating a working model, we may choose a cutoff point d such that $\mathrm{NPV}(d) = \gamma$, an acceptable large value, and then make inferences about $\mathrm{PPV}(d)$. However, since $\hat{\mathrm{NPV}}(c)$ may not estimate $\mathrm{NPV}(c)$ well when c is near 0 or 1, the above cutoff point d may not be stable for the finite sample case. Moreover, even when $g(\beta_0' Z)$ is continuous, empirically $\mathrm{NPV}(c)$ may not be able to reach a pre-specified γ. Therefore, for the class of classification rules $I(g(\hat{\beta}' Z) > c)$, we choose the cutoff point $\hat{c}$ such that $\hat{\mathrm{SE}}(\hat{c}) = \gamma$, an acceptable level of sensitivity, as we did in Section 3. We then compute the corresponding $\hat{\mathrm{PPV}}(\hat{c})$ and $\hat{\mathrm{NPV}}(\hat{c})$. For example, for Model II with $\gamma = .69$, $\hat{c} = .23$ and (4.1) and (4.2) are .35 and .77, respectively. On the other hand, for Model III with the same γ, $\hat{c} = .29$ and (4.1) and (4.2) are .54 and .85.

To construct confidence intervals for $\mathrm{PPV}(c^\dagger)$ and $\mathrm{NPV}(c^\dagger)$, where $\mathrm{SE}(c^\dagger) = \gamma$, one may use the perturbation-resampling scheme to obtain the perturbed versions of $\hat{\mathrm{PPV}}(\hat{c})$ and $\hat{\mathrm{NPV}}(\hat{c})$. Specifically, first, let c^*be the solution of $\mathrm{SE}^*(c^*) = \gamma$, as we did in Section 3. Then, let

$$\mathrm{PPV}^*(c^*) = \frac{\sum_{i=1}^n w_i \{G^*(t \wedge X_i)\}^{-1} I(g(Z_i'\beta^*) \geq c^*, X_i < t) V_i}{\sum_{i=1}^n I(g(Z_i'\beta^*) \geq c^*)) V_i}, \tag{4.3}$$

and

$$\mathrm{NPV}^*(c^*) = \frac{\sum_{i=1}^n w_i \{G^*(t \wedge X_i)\}^{-1} I(g(Z_i'\beta^*) < c^*, X_i \geq t) V_i}{\sum_{i=1}^n I(g(Z_i'\beta^*) < c^*)) V_i}, \tag{4.4}$$

respectively.

It follows from the same argument **used for the OMR estimators** that for large n, the joint distribution of $n^{1/2}(\hat{\text{PPV}}(\hat{c}) - \text{PPV}(c^\dagger))$ and $n^{1/2}(\hat{\text{NPV}}(\hat{c}) - \text{NPV}(c^\dagger))$ can be approximated well by the conditional joint distribution of $n^{1/2}(\text{PPV}^*(c^*) - \hat{\text{PPV}}(\hat{c}))$ and $n^{1/2}(\text{PPV}^*(c^*) - \hat{\text{PPV}}(\hat{c}))$.

For the gene-expression example, for Model II with $\gamma = .69$, 95% confidence intervals for $\text{PPV}(c^\dagger)$ and $\text{NPV}(c^\dagger)$ are (.28, .46) and (.62, .80), respectively. For Model III, the corresponding intervals are (.35, .62) and (.78, .89). Furthermore, for the differences of $\text{PPV}(c^\dagger)$ and $\text{NPV}(c^\dagger)$ between these two models, 95% intervals are (.01, .24) and (.06, .19), respectively.

Note that one can obtain the cross validation counterparts of (4.1) and (4.2) and their distributions can be approximated via (4.3) and (4.4) as we did for the apparent error estimates.

5. REMARKS

Without censoring, an alternative way to evaluate prediction rules may be based on the absolute value of the difference between the future T^0 and its predicted value via a fitted model of Z^0 (Tian et al. 2005). Unfortunately in the presence of censoring, when the support of the censoring is significantly shorter than that of the survival time, it seems rather difficult if not impossible to estimate the mean of the above distance measure well (Sinisi and van der Laan 2005). Furthermore, such a distance measure summarizes the average accuracy of the prediction rules across all time points and does not differentiate the accuracy for classifying short term survivors from that for classifying long term survivors. On the other hand, our procedure is flexible for evaluating classification rules for predicting survivors at any reasonable time point t of interest.

It is important to note that either with or without censored observations involved, our

estimator $\hat{\beta}$ converges to the same value β_0, a root to $u(\beta) = EZ\{I(T < t) - g(\beta_0' Z)\}$, and $\hat{D}(c)$ converges to the same $D(c) = E|I(T < t) - g(\hat{\beta}_0' Z)|$. Therefore, at least for the large sample case, the nuisance censoring distribution of the study does not contaminate the development and evaluation process of the prediction rules. The proposed procedure, however, does require the assumption that the censoring variable is either free of the covariates or its conditional distribution can be estimated consistently using semi-parametric or non-parametric methods when some of the covariates are continuous. If the covariate vector is discrete, a purely non-parametric estimator for the covariate specific censoring distribution can be easily constructed and our procedures can be generalized easily to incorporate the covariate-dependent censoring. Note that even if we let $\hat{\beta}$ be the standard estimator for a commonly used survival model, which does not involve an estimate $\hat{G}(\cdot)$ for the censoring distribution, it is not clear how to construct a consistent estimator of, for example, $D(c)$, with censored observations. Moreover, when the fitted model may not be correctly specified, it is a rather challenging, if not impossible, task to generalize our procedures to handle the covariate-dependent censoring case under a truly nonparametric setting.

In this article, we show how to obtain interval estimates for various prediction precision measures to evaluate prediction rules constructed from censored regression models. Based on the results of our numerical studies, we recommend the interval estimator, which is centered around the cross validate point estimate, for practical usage.

APPENDIX A: CONSISTENCY OF $\hat{c}$ AND $\hat{D}(\hat{c})$

To show that $\hat{c}$ is a consistent estimator of c_0, it suffices to show that $\hat{D}(c)$ converges to $D(c)$, uniformly in c, and $D(c)$ has a unique minimiser c_0 (Newey and McFadden 1994,

Theorem 2.1). To show the uniform consistency of $\hat{D}(c)$, we let

$$\hat{D}(c,\beta) = n^{-1}\sum_{i=1}^{n}\frac{w_i}{\hat{G}(X_i \wedge t)}|I(X_i < t) - I(g(\beta' Z_i) \geq c)|,$$

and $D(c,\beta) = E|I(T^0 < t) - I(g(\beta' Z_i) \geq c)|$. Then it follows from the uniform consistency of $\hat{G}(\cdot)$ (Kalbfleish and Prentice 2002) and a uniform law of large numbers (Pollard 1990) that $\sup_{c,\beta\in\Omega}|\hat{D}(c,\beta) - D(c,\beta)| \to 0$ almost surely, where Ω is the compact parameter space for β around β_0. This, coupled with the fact that $\hat{\beta}$ converges to β_0, implies that $\hat{D}(c) = \hat{D}(c,\hat{\beta})$ is uniformly consistent for $D(c) = D(c,\beta_0)$. Now, to show that $D(c)$ has a unique minimizer, we write

$$D(c) = \text{pr}(T \geq t) + E[\{2I(T < t) - 1\}I(g(\beta_0' Z) < c)]$$

$$= \text{pr}(T \geq t) + E[\{2h_0(g(\beta_0' Z)) - 1\}I(g(\beta_0' Z) < c)] = \text{pr}(T \geq t) + \int_0^{F_0(c)}\{2h_0(F_0^{-1}(x)) - 1\}dx,$$

where $F_0(y) = P(g(\beta_0' Z) < y)$ and $h_0(y) = P(T < t \mid g(\beta_0' Z) = y)$. Thus, assuming that $F_0(y)$ is strictly increasing, $D(c)$ has a unique minimizer if and only if

$$\zeta(u) = \int_0^u \{2h_0(F_0^{-1}(x)) - 1\}dx = 2\int_0^u h_0(F_0^{-1}(x))dx - u$$

has a unique minimizer which is guaranteed if $h_0(\cdot)$ is an increasing function. This concludes that $\hat{c}$ is a consistent estimator of c_0. The consistency of $\hat{D}(\hat{c})$ follows directly from the consistency of $\hat{c}$ and the uniform convergence of $\hat{D}(c)$ to $D(c)$.

APPENDIX B: LARGE DISTRIBUTION OF $W = n^{1/2}(\hat{D}(\hat{c}) - D_0)$

To derive the limiting distribution of W, we let $W(c,\beta) = n^{1/2}\{\hat{D}(c,\beta) - D(c,\beta)\}$ and note that

$$W = W(\hat{c},\hat{\beta}) + n^{1/2}\{D(\hat{c},\hat{\beta}) - D(\hat{c},\beta_0)\} + n^{1/2}\{D(\hat{c}) - D_0\}. \tag{B.1}$$

We first derive the large sample distribution for $W(c,\beta)$. To this end, we note that

$$\hat{W}_G(t) = \frac{n^{1/2}\{G(t) - \hat{G}(t)\}}{G(t)} \simeq n^{-1/2}\sum_{i=1}^{n}\psi_i(t),$$

and $\hat{W}_G(t)$ converges weakly to a zero-mean Gaussian process indexed by t (Kalbfleish and Prentice 2002), where $\psi_i(t) = \int_0^t dM_i(u)/\pi_X(u)$, $\pi_X(t) = \mathrm{pr}(X_i \geq t)$, $M_i(t) = I(X_i \leq t, \delta_i = 0) - \int_0^t I(X_i \geq u)d\Lambda_C(u)$, and $\Lambda_C(\cdot)$ is the cumulative hazard function for the common censoring variable. This, together with a uniform law of large numbers and Lemma A.1 of Billias, Gu and Ying (1997), implies that

$$W(c,\beta) \approx n^{-1/2}\sum_{i=1}^{n}\{D_i(c,\beta) - D(c,\beta)\} + \int_0^t \hat{W}_G(s)d\hat{\gamma}(s;\beta) \approx n^{-1/2}\sum_{i=1}^{n} W_{1i}(c,\beta), \quad \text{(B.2)}$$

where $D_i(c,\beta) = w_i|I(T_i < t) - I(\beta' Z_i \geq c)|/G(T_i \wedge t)$, $\hat{\gamma}(s;\beta) = n^{-1}\sum_{i=1}^{n} D_i(c,\beta)I(T_i \wedge t \leq s)$, and $W_{1i}(c,\beta) = D_i(c,\beta) - D(c,\beta) + \int_0^t \psi_i(s)dE\{\hat{\gamma}(s;\beta)\}$. It follows from a functional central limit theorem (Pollard 1990, chap. 10) that $W(c,\beta)$ converges weakly to a zero mean Gaussian process in (c,β) and, thus, $W(\hat{c},\hat{\beta})$ is asymptotically equivalent to $W(c_0,\beta_0)$.

It follows from the consistency of $\hat{c}$ and a Taylor series expansion that the second term in (B.1) is asymptotically equivalent to $n^{1/2}\{D(c_0,\hat{\beta}) - D(c_0,\beta_0)\} \approx \dot{D}_2(c_0,\beta_0)' n^{1/2}(\hat{\beta} - \beta_0)$, where $\dot{D}_2(c,\beta) = \partial D(c,\beta)/\partial\beta$. Now, by a taylor series expansion of $U(\beta)$ around β_0 and the uniform consistency of $\hat{G}(\cdot)$, we have $n^{1/2}(\hat{\beta} - \beta_0) \approx A(\beta_0)n^{1/2}U(\beta_0)$, where $A(\beta) = -\{\partial u(\beta)/\partial\beta\}^{-1}$. This implies that

$$n^{1/2}(\hat{\beta} - \beta_0) \approx A(\beta_0)\left\{n^{-1/2}\sum_{i=1}^{n} e_i(\beta_0) + \int_0^t \hat{W}_G(s)d\hat{K}(s;\beta_0)\right\} \approx n^{-1/2}\sum_{i=1}^{n} W_{2i}(\beta_0),$$

where $e_i(\beta) = w_i Z_i\{I(T_i < t) - g(\beta' Z_i)\}/G(T_i \wedge t)$, $\hat{K}(s;\beta) = n^{-1}\sum_{i=1}^{n} e_i(\beta)I(T_i \wedge t \leq s)$, and $W_{2i}(\beta) = A(\beta)\{e_i(\beta) + \int_0^t \psi_i(s)dE\{\hat{K}(s;\beta)\}$. Therefore

$$n^{1/2}\{D(\hat{c},\hat{\beta}) - D(\hat{c},\beta_0)\} \approx n^{-1/2}\sum_{i=1}^{n}\dot{D}_2(c_0,\beta_0)' W_{2i}(\beta_0), \quad \text{(B.3)}$$

The weak convergence of the process $W(c,\beta)$ and the convergence of $n^{1/2}(\hat{\beta}-\beta_0)$ imply that the process $n^{1/2}\{\hat{D}(c)-D(c)\} = W(c,\hat{\beta}) + n^{1/2}\{D(c,\hat{\beta})-D(c)\}$ is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}\{W_{1i}(c,\beta_0) + \dot{D}_2(c,\beta_0)'W_{2i}(\beta_0)\}$ and is tight in c. Now, since $0 \geq n^{1/2}\{D(\hat{c}) - D(c_0)\} = n^{1/2}\{\hat{D}(\hat{c}) - \hat{D}(c_0)\} - n^{1/2}\{\hat{D}(\hat{c}) - D(\hat{c}) - \hat{D}(c_0) + D(c_0)\} \geq -n^{1/2}\{\hat{D}(\hat{c}) - D(\hat{c}) - \hat{D}(c_0) + D(c_0)\}$, $|n^{1/2}\{D(\hat{c}) - D(c_0)\}| \leq |n^{1/2}\{\hat{D}(\hat{c}) - D(\hat{c}) - \hat{D}(c_0) + D(c_0)\}|$. This, together with the tightness of the process $n^{1/2}\{\hat{D}(c) - D(c)\}$, implies that $n^{1/2}\{D(\hat{c}) - D(c_0)\} = o_p(1)$. Note that, when Z is discrete, it is straightforward to show that $|n^{1/2}\{\hat{D}(\hat{c}) - D(\hat{c}) - \hat{D}(c_0) + D(c_0)\}| = o_p(1)$ since $\mathrm{pr}(\hat{c} = c_0) \to 1$. It then follows from (B.2) and (B.3) that

$$W \approx n^{-1/2}\sum_{i=1}^{n}\{W_{1i}(c_0,\beta_0) + \dot{D}_2(c_0,\beta_0)'W_{2i}(\beta_0)\}.$$

By the central limit theorem, W converges in distribution to a normal with mean 0 and variance $E[(W_{1i}(c_0,\beta_0) + W_{2i}(\beta_0))^2]$

APPENDIX C: LARGE SAMPLE DISTRIBUTION OF $n^{1/2}\{\hat{\mathcal{D}}(\hat{c}_v) - D_0\}$ AND $n^{1/2}\{\hat{\mathbb{D}}(\hat{c}_{rv}) - D_0\}$

Let $\{\xi_i; i = 1,\cdots,n\}$ be n exchangeable discrete random variables uniformly distributed over $\{1,2,\cdots,K\}$, independent of the data, and satisfy that $\sum_{i=1}^{n} I(\xi_i = k) = n/K, k = 1,\cdots,K$. Let $\hat{D}_{(k)}(c,\beta)$ denote $\hat{D}(c,\beta)$ evaluated based observations in $\mathcal{I}_k$, then $\hat{D}_{(k)}(c) = \hat{D}_{(k)}(c,\beta_{(-k)})$. Then for the kth partition, we have

$$\hat{D}_{(k)}(\hat{c}_v) - D_0 = \hat{D}_{(k)}(\hat{c}_v,\hat{\beta}_{(-k)}) - D(\hat{c}_v,\hat{\beta}_{(-k)}) + D(\hat{c}_v,\hat{\beta}_{(-k)}) - D(\hat{c}_v,\beta_0) + D(\hat{c}_v,\beta_0) - D_0.$$

It follows from the same argument as given in Appendix B that

$$\hat{\beta}_{(-k)} - \beta_0 = \frac{K}{n(K-1)} \sum_{i=1}^{n} I(\xi_i \neq k) W_{2i}(\beta_0) + o_p(n^{-1/2}) \tag{C.1}$$

$$\hat{D}_{(k)}(\hat{c}_v, \hat{\beta}_{(-k)}) - D(\hat{c}_v, \hat{\beta}_{(-k)}) = n^{-1} \sum_{i=1}^{n} I(\xi_i = k) W_{1i}(c_0, \beta_0) + o_p(n^{-1/2}), \tag{C.2}$$

$$D(\hat{c}_v, \hat{\beta}_{(-k)}) - D(\hat{c}_v, \beta_0) = \frac{1}{n(K-1)} \dot{D}_2(c_0, \beta_0)' \sum_{i=1}^{n} I(\xi_i \neq k) W_{2i}(\beta_0) + o_p(n^{-1/2}), \tag{C.3}$$

and $D(\hat{c}_v, \beta_0) - D_0 = o_p(n^{-1/2})$, where the p is the product probability measure generate by that of $\{\xi_1, \cdots, \xi_n\}$ and the data. Therefore

$$\hat{D}_{(k)}(\hat{c}_v) - D_0 = n^{-1} \sum_{i=1}^{n} \left\{ I(\xi_i = k) W_{1i}(c_0, \beta_0) + I(\xi_i \neq k) \frac{1}{K-1} \dot{D}_2(c_0, \beta_0)' W_{2i}(\beta_0) \right\}.$$

It follows that

$$\hat{\mathcal{D}}(\hat{c}_v) - D_0 = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ I(\xi_i = k) W_{1i}(c_0, \beta_0) + I(\xi_i \neq k) \frac{1}{K-1} \dot{D}_2(c_0, \beta_0)' W_{2i}(\beta_0) \right\}.$$

Now, since $\sum_{k=1}^{K} I(\xi_i = k) = 1$ and $\sum_{k=1}^{K} I(\xi_i \neq k) = K - 1$, it is straightforward to show that

$$\hat{\mathcal{W}} = n^{1/2} \left\{ \hat{\mathcal{D}}(\hat{c}_v) - D_0 \right\} = n^{-1/2} \sum_{i=1}^{n} \left\{ W_{1i}(c_0, \beta_0) + \dot{D}_2(c_0, \beta_0)' W_{2i}(\beta_0) \right\} + o_p(1).$$

Thus $\hat{\mathcal{W}}$ is asymptotically equivalent to W.

For the general cross validation procedure, without loss of generality, we assume $n/n_v = K$, then $n^{1/2}(\hat{\mathbb{D}}(\hat{c}_{rv}) - D_0) = E_\xi[n^{1/2}\{\hat{\mathcal{D}}(\hat{c}_v) - D_0\}]$, where the expectation is with respect to random variables $\{\xi_1, \cdots, \xi_n\}$. It follows from Theorem 3.1 of Chatterjee and Bose (2005), the approximations given in (C.1), (C.2) and (C.3) that $n^{1/2}\{\hat{\mathcal{D}}(\hat{c}_v) - D_0\} = W + o_{p^*}(1)$, where p^* is the product probability measure generated by that of $\{\xi_1, \cdots, \xi_n\}$ and the data. Consequently, by a Markov inequality, $\text{pr}(|E_\xi[n^{1/2}\{\hat{\mathcal{D}}(\hat{c}_v) - D_0\}] - W| > \epsilon) \leq \text{pr}(E_\xi|n^{1/2}\{\hat{\mathcal{D}}(\hat{c}_v) - D_0\} - W| >$

$\epsilon) \leq \epsilon^{-1} E^* |n^{1/2}\{\hat{\mathcal{D}}(\hat{c}_v) - D_0\} - W| \to 0$, for any $\epsilon > 0$, where the last expectation E^* is with respect to both $\{\xi_1, \cdots, \xi_n\}$ and the data. It follows that $n^{1/2}(\hat{\mathbb{D}}(\hat{c}_{rv}) - D_0)$ is asymptotically equivalent to W.

References

Brier, G. W. (1950), "Verification of Forecasts Expressed in terms of Probability," *Monthly Weather Review,* 78, 1–3.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees,* New York: Chapman & Hall.

Bilias, Y., Gu, M. and Ying, Z. (1997), "Towards a General Asymptotic Theory for Cox Model with Staggered Entry," *The Annals of Statistics,* 25, 662–682.

Cai, T., Tian, L. and Wei, L. J. (2005), "Semiparametric Box-Cox Power Transformation Models for Censored Survival Observations," *Biometrika,* 92, 619–632.

Chatterjee, S. and Bose, A. (2005), "Generalized Bootstrap for Estimating Equations," *The Annals of Statistics,* 33, 414–436.

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 Tables" (with Discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 187–220.

Dickson, E., Fleming, T., Grambsch, P., Fisher, L., and Langworthy, A. (1989), "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making," *Hepatology,* 10, 1–7.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and their Application,* Cambridge: Cambridge University Press.

Efron, B. (1982), *The Jackknife, The Bootstrap, and Other Resampling Plans,* SIAM NSF-CBMS, Monograph # 38.

——(1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-

Validation," *Journal of the American Statistical Association,* 78, 316–331.

——(1986), "How Biased is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association,* 81, 461–470.

——(2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of the American Statistical Association,* 99, 619–632.

Efron, B. and Tibshirani, R. (1997), "Improvements on Cross-Validation: The .632 + Bootstrap Method," *Journal of the American Statistical Association,* 92, 548–560.

Fleming, T. R. and Harrington, D. P. (1991), *Counting Processes and Survival Analysis,* New York: John Wiley & Sons, Inc.

Heagerty P. J., Lumley, T., and Pepe, M. S. (2000), "Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker," *Biometrics,* 56, 337–344.

Heagerty P. J. and Zheng, Y. (2005), "Survival Model Predictive Accuracy and ROC Curves," *Biometrics,* 61, 92–105.

Kalbfleish, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data* (2nd ed.), New York: John Wiley & Sons, Inc.

Korn, E and Simon, R. (1990), "Measures of Explained Variation for Survival Data," *Statistics in Medicine,* 9, 487–503.

Mallows, C. L. (1973), "Some Comments on C_P," *Technometrics,* 15, 661–675.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition,* New York: John Wiley & Sons, Inc.

Mittlbock, M. and Schemper, M. (1996), "Explained Variation for Logistic Regression," *Statistics in Medicine,* 15, 1987–1997.

Moskowitz, C. S. and Pepe, M. S. (2004), "Quantifying and Comparing the Accuracy of

Binary Biomarkers when Predicting a Failure Time Outcome," *Statistics in Medicine,* 23, 1555–1570.

Newey, W. K. and McFadden, D. L. (1994), "Large Sample Estimation and Hypothesis Testing," In *Handbook of Econometrics*(Vol. 4), eds. R. F. Engle and D. L. McFadden, Amsterdam: Elsevier B. V.

O'Quigley, J. and Xu, R. (2001), "Explained Variation in Proportional Hazards Regression," In *Handbook of Statistics in Clinical Oncology,* ed. J. Crowley, New York: Marcel Dekker, pp. 397–409.

Park, Y. and Wei, L. J. (2003), "Estimating Subject-Specific Survival Functions under the Accelerated Failure Time Model," *Biometrika,* 90, 717–723.

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction,* Oxford: Oxford University Press.

Pollard, D. (1990), *Empirical Processes: Theory and Applications,* Hayward, CA: Institute of Mathematical Statistics.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks,* Cambridge: Cambridge University Press.

Shao, J. (1996), "Bootstrap Model Selection," *Journal of the American Statistical Association,* 91, 655–665.

Sinisi, S. E. and van der Laan, M. J. (2005), "Cross-Validated Bagged Prediction of Survival," *Technical Report No. 189, Division of Biostatistics, UC Berkeley, 2005.* http://www.bepress.com/ucbbiostat/paper189/

Spiegelhalter, D. J. (1986), "Probabilistic Prediction in Patient Management and Clinical Trials," *Statistics in Medicine,* 5, 412–433.

Tian, L., Cai, T., Goetghebeur, E. and Wei, L.J. (2005), "Model Evaluation Based on the Distribution of Estimated Absolute Prediction Error." *Harvard University Biostatistics Working Paper Series,* http://www.bepress.com/harvardbiostat/paper35

Tibshirani, R. and Knight, K. (1999), "Model Search by Bootstrap "bumping"," *Journal of Computational and Graphical Statistics,* 8, 671–686.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards, R. (2002), "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *The New England Journal of Medicine,* 347, 1999–2009.

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.

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association,* 93, 120–131.

Zheng, Y, Cai, T. and Feng, Z. (2006), Application of the Time-Dependent ROC Curves for Prognostic Accuracy with Multiple Biomarkers," *Biometrics,* 62, 279–287.

Zhou, X.H., Obuchowski, N. A. and McClish, D. K. (2002), *Statistical Methods in Diagnostic Medicine,* New York: John Wiley & Sons, Inc.

Table 1. *Comparing Various Model-based Prediction Rules for 10-Year Survivors with Mayo Biliary Cirrhosis Data*

Model[1]	Apparent Error	10-fold CV	Random CV	
	$\hat{D}(\hat{c})$ (s.e.)[2]	$\hat{\mathcal{D}}(\hat{c}_v)$	$\hat{\mathbb{D}}(\hat{c}_{rv})$	CI for difference[3]
I	.30 (.050)	.30	.34	$(.03, .21)$[4]
II	.16 (.042)	.18	.22	$(-.03, .05)$[5]
III	.16 (.043)	.18	.21	$(-.07, .06)$[6]
IV	.17 (.038)	.18	.21	

(1) Model I: g(intercept+age);
Model II: g(intercept+age+log(bilirubin)+log(albumin)+ edema + log(protime));
Model III: g(intercept+age+log(bilirubin)+log(albumin);
Model IV: g(intercept+age+log(bilirubin)), where
$g(y) = 1 - \exp\{-\exp(y)\}$. (2) s.e.: estimated standard error. (3) 95% confidence interval for the difference of the OMRs of two competing models.
(4) Model I- Model II; (5) Model II - Model III; (6) Model III - Model IV.

Table 2. *Comparing Various Model-based Prediction Rules for 10-Year Survivors with the Breast Cancer Data*

Model[1]	Apparent Error	10-fold CV	Random CV
	$\hat{D}(\hat{c})$ (s.e.)[2]	$\hat{\mathcal{D}}(\hat{c}_v)$	$\hat{\mathbb{D}}(\hat{c}_{rv})$
I	.30 (.031)	.29	.30
II	.28 (.033)	.30	.28
III	.25 (.036)	.27	.28
van de Vijver[3]	.35 (.050)	–	–

(1) Model I: g(intercept); Model II: g(intercept+Node+ER);
Model III: g(intercept+Node+ER+Gene), where $g(y) = 1 - \exp\{-\exp(y)\}$.
(2) s.e.: estimated standard error. (3) Based on the classification rule in van de Vijver et al.(2002).

Table 3. *Empirical bias, root mean square error (RMSE) and coverage probability based on apparent error (AE), 10-fold cross-validation (CV_{10}), random cross-validation ($CV_{1/3}$) with sample size 300 and $t = 10$ years*

		Observed censoring			No censoring		
	Model*	AE	CV_{10}	$CV_{1/3}$	AE	CV_{10}	$CV_{1/3}$
Bias	I	−.038	−.016	.004	−.015	−.007	.002
	II	−.036	−.021	−.003	−.014	−.009	−.001
	III	−.039	−.016	.006	−.015	−.007	.003
	IV	−.038	−.016	.005	−.015	−.007	.002
	V	−.039	−.015	.006	−.015	−.007	.003
	VI	−.053	−.003	.025	−.020	−.002	.010
RMSE	I	.054	.044	.042	.027	.025	.024
	II	.053	.045	.041	.027	.025	.024
	III	.055	.045	.043	.028	.025	.025
	IV	.054	.044	.042	.027	.025	.024
	V	.054	.045	.043	.028	.025	.024
	VI	.065	.044	.050	.030	.024	.026
Coverage level	I	.887	.939	.962	.926	.949	.958
	II	.912	.944	.968	.932	.945	.962
	III	.882	.935	.962	.919	.941	.947
	IV	.888	.945	.959	.925	.947	.959
	V	.884	.938	.955	.929	.944	.954
	VI	.773	.926	.914	.884	.938	.923

* Model I (true): $D_0 = .262$; Model II (covariate omission): $D_0 = .271$; Model III (wrong functional form): $D_0 = .268$; Model IV (wrong link function): $D_0 = .262$; Model V (wrong link and wrong functional form): $D_0 = .267$. Model VI (over fitting): $D_0 = .262$;

Figure 1. The ROC curves of various prediction models for 10-Year Survivors with the Breast Cancer Data

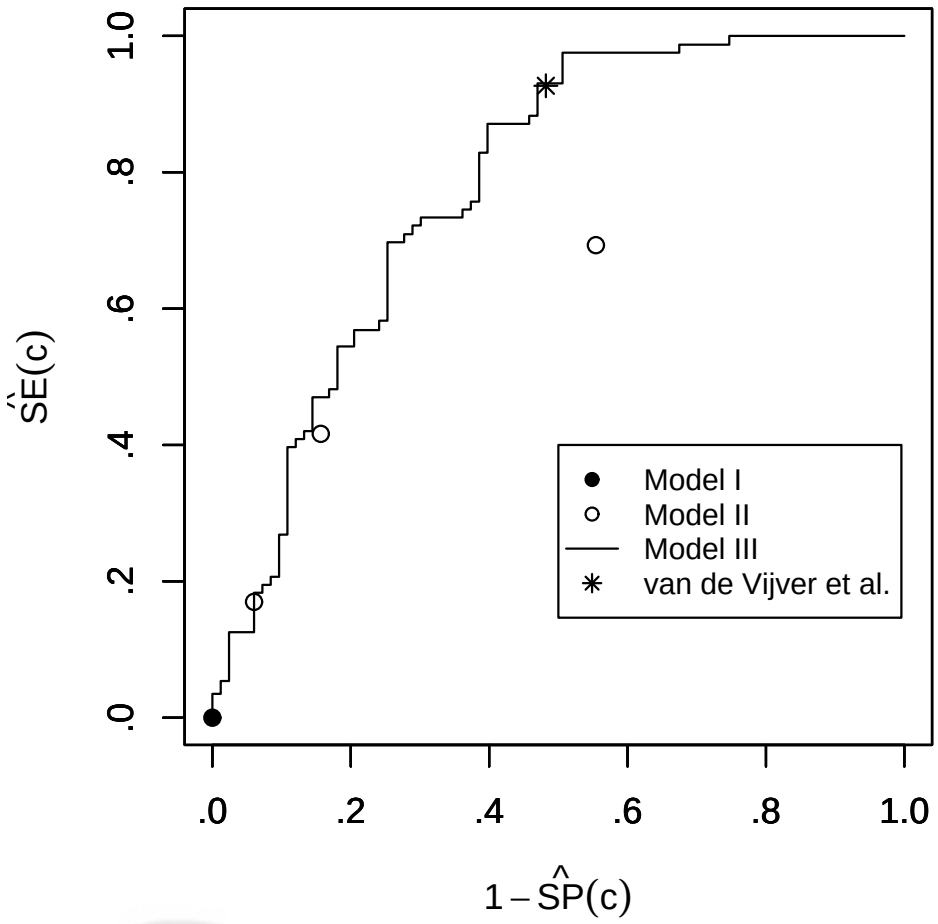

Figure 2. The PPV-NPV curves of various prediction models for 10-Year Survivors with the Breast Cancer Data