Model-free scoring system for risk prediction with application to hepatocellular carcinoma study

Weining Shen[1], Jing Ning[2,*], Ying Yuan[2], Anna S. Lok[3], and Ziding Feng[2]

[1]Department of Statistics, University of California, Irvine, CA, U.S.A

[2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.

[3]Division of Gastroenterology, University of Michigan Health System, U.S.A.

*email: jning@mdanderson.org

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/biom.12750

Biometrics 000, 000–000 DOI: 000
000 0000

Model-free scoring system for risk prediction with application to hepatocellular carcinoma study

Weining Shen[1], Jing Ning[2,*], Ying Yuan[2], Anna S. Lok[3], and Ziding Feng[2]

[1]Department of Statistics, University of California, Irvine, CA, U.S.A

[2]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.

[3]Division of Gastroenterology, University of Michigan Health System, U.S.A.

email: jning@mdanderson.org

Summary: There is an increasing need to construct a risk-prediction scoring system for survival data and identify important risk factors (e.g., biomarkers) for patient screening and treatment recommendation. However, most existing methodologies either rely on strong model assumptions (e.g., proportional hazards) or only handle binary outcomes. In this paper, we propose a flexible method that simultaneously selects important risk factors and identifies the optimal linear combination of risk factors by maximizing a pseudo-likelihood function based on the time-dependent area under the receiver operating characteristic curve. Our method is particularly useful for risk evaluation and recommendation of optimal subsequent treatments. We show that the proposed method has desirable theoretical properties, including asymptotic normality and the oracle property after variable selection. Numerical performance is evaluated on several simulation data sets and an application to hepatocellular carcinoma data.

Key words: Biomarker; Liver cancer; Risk prediction; Scoring system; Time-dependent AUC; Variable selection

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Genomic medical research has generated a large number of candidate biomarkers that have potential use in the early-phase detection and prognosis of many diseases. Compared to the conventional approach based on single biomarker, simultaneously using multiple biomarkers can substantially improve the sensitivity and accuracy of early detection of diseases (Sidransky, 2002; Etzioni et al., 2003). Multiple biomarker-based scoring systems, such as the International Prognostic Scoring System (IPSS) (Greenberg et al., 1997), WHO Prognostic Scoring System (WPSS), and Revised International Prognostic Scoring System (IPSS-R) (Greenberg et al., 2012), have played fundamental roles in the treatment decision-making process. For example, multiple biomarkers have been used to guide treatment decisions for Myelodysplastic syndromes (MDS), a heterogeneous group of myeloid disorders. IPSS or IPSS-R can play a crucial role in differentiating between patients at high risk of disease progression, for whom a more aggressive treatment may be justified, and patients with a minor risk of disease progression, for whom a more conservative treatment may be preferable.

Although developing scoring systems for risk prediction has been an active research area, the vast majority of them have focused on binary outcomes (e.g., developing disease or not). Su and Liu (1993) considered linear discriminant analysis by maximizing the area under the receiver operating characteristic (ROC) curve (AUC). That maximization/classification idea has been extended in various ways, including non-normal distributions (Pepe and Thompson, 2000), generalized linear models (Pepe et al., 2006), maximizing ROC values at specified point (McIntosh and Pepe, 2002), maximizing sensitivity over a range of specificity (Liu et al., 2005), maximizing the empirical AUC (Ma and Huang, 2005), maximizing an ROC-type measure given a continuous gold reference available (Chang, 2013). Yuan and Ghosh (2008) proposed a model-combining algorithm that builds on logistic regression models. Recently, Chen et al. (2015) proposed an empirical-likelihood-based approach to estimate

the confidence intervals of the AUC and find the optimal linear combination of biomarkers. Chen et al. (2016) discussed a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests.

Scoring systems for time-to-event data are increasingly needed in practice. The aforementioned methods for binary outcomes may not be efficient since the observed information from the censored subjects cannot be fully utilized. The existing approaches for constructing scoring systems for time-to-event data mostly rely on the Cox model (Greenberg et al., 1997; D'Avanzo, A., et al., 2004; Kadalayil, L., et al., 2013) or the proportional odds model (Zheng et al., 2006), which may lead to poor performance in risk prediction if the assumption of proportional hazards (odds) is violated. Hence, it is desirable to develop an approach that is robust to model mis-specification for time-to-event data. In addition, given that the recent advance of biomedical research has produced a large volume of candidate biomarkers that may be useful for risk prediction, it is crucial that the proposed approach can also perform variable selection to increase the efficiency and interpretability of the resulting scoring systems. A simple pre-selection procedure based on unjustified criteria/incorrect models (e.g., marginal correlation) may lead to an undesirable subset of risk factors.

The objective of this paper is to fill these gaps by providing a method that simultaneously identifies useful risk factors and constructs optimal risk scores for time-to-event outcomes. The rest of the paper is organized as follows. We discuss a motivating example in Section 2. In Section 3, we construct a pseudo-likelihood function and solve the corresponding estimation equations to optimize the incidence/dynamic time-dependent AUC, which has a close connection to the concordance summaries (Heagerty and Zheng, 2005). In the estimating procedure, the selection of biomarkers proceeds by regularization with adaptive lasso penalty functions (Zou, 2006). Computationally, we consider a kernel-smoothing technique to deal with the non-smooth objective function (Zeng and Lin, 2007). Large-sample properties including $\sqrt{n}$-

consistency and the oracle property after variable selection are derived in Section 4. We present some simulations studies and an application to hepatocellular carcinoma data in Sections 5 and 6. We provide proofs and technical details in the Supplementary file.

2. A motivating example

Hepatocellular carcinoma (HCC) is a primary malignancy of the liver and is now the third leading cause of cancer deaths worldwide, with over 500,000 people affected (Kadalayil, L., et al., 2013). The five-year survival rate of patients with HCC is low ($< 10\%$) (Everhart and Ruhl, 2009), due to late detection and lack of effective treatment options for advanced-stage HCC. There is an urgent need for methods to detect HCC when it is in an early manageable stage of disease that is amenable to curative treatments such as surgical resection, liver transplantation or radiofrequency ablation (Santi, V., et al., 2010).

Most cases of HCC are associated with cirrhosis (80%) of which chronic hepatitis B or C are the most common causes. However, most patients with liver cirrhosis do not develop HCC; only 1-5% progress to HCC annually (Davis, G. L., et al., 2010). Only a small part of this variation in HCC risk can be predicted or explained by the existing knowledge of HCC risk factors. For patients infected with hepatitis C, previous studies have suggested a number of possible risk factors, including the degree of liver fibrosis and cirrhosis, serum biomarkers (e.g., alpha-fetoprotein [AFP], des-gamma-carboxyprothrombin [DCP]), diabetes, obesity, use of tobacco, and excessive alcohol consumption (Lok et al., 2009). Patients that are at high risk will be screened with biomarkers that are positive or abnormal when early stage tumor is present. Among HCC biomarkers, AFP is most commonly used for screening and diagnosis, but its sensitivity and specificity are poor (Zhu, W. W., et al., 2013). Other tumor biomarkers have been proposed to complement or substitute for AFP in HCC detection, such as DCP (Song et al., 2014). Several previous studies have recommended that a combined test of AFP and DCP increases the sensitivity or specificity of early HCC detection over the use

of a single biomarker (Song, P., et al., 2013). For example, in a study of 210 Chinese patients, a combined test of DCP and AFP had a sensitivity of 78.3%, which was higher than that of DCP alone (53.3%) or AFP alone (58.3%) (Cui, R., et al., 2003). However, using multiple markers increase sensitivity but often decrease specificity (Lok et al., 2010).

The precise role of these risk factors in the prediction of HCC is not yet known. There is thus a great need for reliable statistical models that can efficiently combine risk factors for prediction purposes. Some efforts in that direction have been made. Sanyal, A. J., et al. (2006) proposed an HCC risk prediction method based on logistic regression. Lok et al. (2009) proposed an approach based on the proportional hazard model. These approaches, however, rely on strong model assumptions (i.e., logistic model or proportional hazards), and their performance may be compromised when the assumptions are violated. In this paper, motivated by a data set recently collected from a randomized two-arm trial over 10 sites in the U.S., we propose new statistical methods that identify risk factors associated with the development of HCC and construct a score formula for risk prediction without any model assumptions on the time to the development of HCC.

3. Method

3.1 *Notation*

Let T be the time measured from the onset of the initial event to the failure event, referred as the survival time. Denote $\boldsymbol{X}$ to be a p-vector of the covariates including the biomarkers and the other patients' characteristics. Given covariates $\boldsymbol{X}$, define $S(\boldsymbol{X})$ as a scoring system, where higher scores are related to higher risk levels and shorter survival times. We assume that the censoring time C is independent of T conditional on $\boldsymbol{X}$. Consider a study cohort with n patients. Let the observed data $(Z_1, \delta_1, \boldsymbol{X}_1), \ldots, (Z_n, \delta_n, \boldsymbol{X}_n)$ be independent and identically distributed copies of $(Z, \delta, \boldsymbol{X})$, where $Z = \min(T, C)$, and $\delta = \mathbb{1}(T \leqslant C)$ is the

censoring indicator. Note that the survival time Z is allowed to depend on the covariates $\boldsymbol{X}$. At a given time point t, define the risk set as $\mathcal{R}(t) = \{j : Z_j > t\}$.

We characterize the performance of the score $S(\boldsymbol{X})$ by the time-dependent incidence/dynamic (I/D) AUC definition (Heagerty and Zheng, 2005), in which the time-dependent incidence sensitivity and dynamic specificity are defined by classifying patients into case and control groups at each time point based on their survival status,

$$\mathrm{TP}_t^{\mathrm{I}}(c) = \mathrm{P}\left\{S(\boldsymbol{X}) > c | T = t\right\}, \quad \mathrm{FP}_t^{\mathrm{D}}(c) = \mathrm{P}\left\{S(\boldsymbol{X}) > c | T > t\right\}.$$

Here $\mathrm{TP}_t^{\mathrm{I}}(c)$ (sensitivity) measures the expected fraction of subjects with a score value greater than c among individuals who experience the event of interest at time t, while $1 - \mathrm{FP}_t^{\mathrm{D}}(c)$ (specificity) measures the fraction of subjects with a score value less than or equal to c among those who survive beyond time t. Then the time-dependent ROC curve is defined by

$$\mathrm{ROC}_t^{\mathrm{I/D}}(p) = \mathrm{TP}_t^{\mathrm{I}}\{(\mathrm{FP}_t^{\mathrm{D}})^{-1}(p)\}, \quad \text{for } p \in [0, 1].$$

and its time-dependent AUC can be obtained as

$$\mathrm{AUC}(t) = \mathrm{P}\left\{S(\boldsymbol{X}_i) > S(\boldsymbol{X}_j) | T_i = t, T_j > t\right\}. \tag{1}$$

3.2 *Estimation*

We model the risk score $S(\boldsymbol{X})$ for the time-to-event outcome by using a smooth function of patients' characteristic, denoted by $S(\boldsymbol{X}; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a finite-dimensional parameter. A commonly used linear model summarizes the patient information as follows,

$$S(\boldsymbol{X}_i; \boldsymbol{\beta}) = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}. \tag{2}$$

For identification purposes, we set $\beta_1 = 1$ in model (2). More generally, we may consider a nonlinear score system,

$$S(\boldsymbol{X}_i; \boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{k=1}^{K} \beta_{jk} \psi_k(X_{ij}), \tag{3}$$

where $\psi_1, \ldots, \psi_K$ are pre-specified nonlinear basis functions, such as polynomials. It is also possible to incorporate the interactions between biomarkers, e.g., $S(\boldsymbol{X}_i; \boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j X_{ij} + \sum_{s=1}^{p} \sum_{t=l+1}^{p} \eta_{st} X_{is} X_{it}$.

Let $t_1 < \cdots < t_M$ be the ordered unique failure times for $\{Z_1, \cdots, Z_n\}, M \leqslant n$. At each time point t_m, the subjects in the risk set $\mathcal{R}(t_m)$ can be divided into two groups:

$$\mathcal{R}^L(t_m) = \{j : S(\boldsymbol{X}_i; \boldsymbol{\beta}) > S(\boldsymbol{X}_j; \boldsymbol{\beta}), Z_j > t_m, Z_i = t_m\},$$
$$\mathcal{R}^H(t_m) = \{j : S(\boldsymbol{X}_i; \boldsymbol{\beta}) \leqslant S(\boldsymbol{X}_j; \boldsymbol{\beta}), Z_j > t_m, Z_i = t_m\}.$$

The first group $\mathcal{R}^L(t_m)$ can be viewed as the set of patients with relatively low risk, whose score values are lower than $S(\boldsymbol{X}_i; \boldsymbol{\beta})$; whereas the second group, $\mathcal{R}^H(z)$, can be viewed as the set of patients with relatively higher risk compared with subject i. By the definition of AUC in (1), it is natural to use the proportional of observing a low-risk patient in the risk set, $|\mathcal{R}^L(t)|/|\mathcal{R}(t)|$, as an estimator of AUC$(t)$, where $|A|$ denotes the size of a set A. In other words, the estimated AUC(t) is the empirical concordance probability $P(S(\boldsymbol{X}_i; \boldsymbol{\beta}) > S(\boldsymbol{X}_j; \boldsymbol{\beta}) | Z_j > Z_i)$. Therefore, we can construct a pseudo-likelihood function by multiplying the empirical concordance probabilities at all uncensored observations i together,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{M} \widehat{\text{AUC}(t_i)} = \prod_{i=1}^{M} \frac{|\mathcal{R}^L(t_i)|}{|\mathcal{R}^L(t_i)| + |\mathcal{R}^H(t_i)|}. \tag{4}$$

Then the estimation of $\boldsymbol{\beta}$ proceeds by maximizing the following log-pseudo-likelihood function, termed the objective function

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \left\{ \frac{\sum_{j=1}^{n} \mathbb{1}(Z_j > Z_i) \mathbb{1}(S(\boldsymbol{X}_i; \boldsymbol{\beta}) - S(\boldsymbol{X}_j; \boldsymbol{\beta}) > 0)}{\sum_{j=1}^{n} \mathbb{1}(Z_j > Z_i)} \right\}. \tag{5}$$

Note that there are several different definitions of time-dependent ROC. We choose to work with the incident/dynamic (I/D) time-dependent ROC over other definitions (e.g., cumulative/dynamic) because it is closely connected with the concordance probability, which allows the construction of the pseudo-likelihood in (5). Moreover, it provides a natural way to define a weighted time-averaged summary of the AUC, which is called IAUC. Heagerty

and Zheng (2005) showed that the IAUC is equivalent to Kendall's τ with a proper weight function.

A major challenge with the maximization of (5) is non-smoothness due to the indicator functions involved. The objective function $\ell_n(\boldsymbol{\beta})$ may remain the same under a small amount of perturbation of $\boldsymbol{\beta}$; hence the finite-sample solution may not be unique. Moreover, because of the non-smoothness property, the maximization is computationally difficult and cannot be solved using standard optimization algorithms designed for continuous functions. One possible solution is to use the Nelder-Mead method by Nelder and Mead (1965), termed the exact method. When there are more than a few covariates, this method can be computationally intensive. Alternatively, we adopt a smoothing kernel to approximate the indicator (Zeng and Lin, 2007). Specifically, we propose to maximize the following approximation to $\ell_n(\boldsymbol{\beta})$:

$$\ell_n^s(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \log\left\{ \frac{\sum_{j=1}^{n} 1\!\!1(Z_j > Z_i) \int_{-\infty}^{S(\boldsymbol{X}_i;\boldsymbol{\beta})-S(\boldsymbol{X}_j;\boldsymbol{\beta})} K(u;h_n)du}{\sum_{j=1}^{n} 1\!\!1(Z_j > Z_i)} \right\}, \tag{6}$$

where $K(\cdot, h_n)$ is a symmetric kernel function with bandwidth h_n converging to 0 as $n \to \infty$. In principle, any smooth symmetric probability density functions can be used here, such as normal, logistic and t-distributions. We choose a standard normal kernel for simplicity and computational tractability. The bandwidth can be chosen by either cross-validation or using the optimal choice of the bandwidth for density estimation problems (Jones, 1990); see Section 4 of Supplementary File for more details.

3.3 *Variable selection*

With rapid developments occurring in biomedical research, a very large number of biomarkers have become available for the construction of the scoring system. However, many biomarkers might not be directly related to the risk. Therefore, it is of great importance to include a variable selection procedure that removes the redundant information and identifies useful biomarkers/risk factors in making clinical decisions. Accordingly, we consider adopting shrinkage penalties for variable selection. We restrict the discussion to the linear score system

(2) for convenience. The nonlinear system can be treated in a similar way. We propose to maximize the following loss function,

$$Q_n(\boldsymbol{\beta}) = \ell_n^s(\boldsymbol{\beta}) - \lambda_n \sum_{j=2}^{p} J(|\beta_j|), \tag{7}$$

where λ_n is a tuning parameter and $J(.)$ is a penalty function. The available choices for $J(.)$ include smoothly clipped absolute deviation (SCAD, Fan and Li, 2001) and adaptive LASSO (Zou, 2006). Here we use an adaptive LASSO penalty, i.e., $J(|\beta_j|) = |\beta_j|/|\tilde{\beta}_j|$, where $\tilde{\beta}_j$ is the j-th element of $\tilde{\boldsymbol{\beta}}$, the solution to $\text{argmax}_{\boldsymbol{\beta}} \ell_n^s(\boldsymbol{\beta})$. This penalty function can be regarded as an asymptotic version of L_0-penalty as long as $\tilde{\boldsymbol{\beta}}$ is consistent (Zhang and Lu, 2007).

Computationally, we adopt the coordinate descent algorithm (Friedman et al., 2007) that solves the optimization problem by updating one parameter at one time while keeping all others fixed. We first fix an initial estimate $\tilde{\boldsymbol{\beta}}$ by solving $\text{argmax}_{\boldsymbol{\beta}} \ell_n^s(\boldsymbol{\beta})$. We then update β_k $(k = 2, \ldots, p)$ by solving the one-dimensional optimization

$$\text{argmin}_{\beta_k} \lambda_n \frac{|\beta_k|}{|\tilde{\beta}_k|} - \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \left\{ \frac{\sum_{j=1}^{n} \mathbb{1}(Z_j > Z_i) \int_{-\infty}^{h_n^{-1}(\beta_k(X_{ki} - X_{kj}) + c_{ij})} K(u) du}{\sum_{j=1}^{n} \mathbb{1}(Z_j > Z_i)} \right\},$$

keeping all other parameters fixed. Here, c_{ij} is a constant that depends on indexes i and j only. The iterative procedures continue until the pre-specified convergence criteria is met. We denote the final solution by $\hat{\boldsymbol{\beta}}$. The selection of λ_n can be achieved by cross-validation based on the prediction performance. We will discuss this issue in details in Section 5.2.

4. Asymptotic results

In this section, we establish asymptotic properties of the proposed estimators. Denote the true value of $\boldsymbol{\beta}$ in the scoring system by $\boldsymbol{\beta_0}$. We first show that the kernel-based smoothing estimator is consistent and asymptotically normally distributed without the variable selection procedure, by using kernel approximation theory and the techniques of Zeng and Lin (2007).

THEOREM 1: *Under Conditions (C1) and (C2) listed in the Supplementary file,* $\tilde{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$

almost surely and $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \to_d N(\mathbf{0}, \boldsymbol{V_1})$*, in which* $\boldsymbol{V_1} = \boldsymbol{\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}}$*, and*

$$\boldsymbol{\Sigma_1} = E\{-\nabla_{\boldsymbol{\beta}} w_n^i(\boldsymbol{\beta}); \boldsymbol{\beta_0}\}, \quad \boldsymbol{\Sigma_2} = E\{w_n^i(\boldsymbol{\beta})^{\otimes 2}; \boldsymbol{\beta_0}\},$$

$$w_n^i(\boldsymbol{\beta}) = \frac{\frac{1}{nh_n}\sum_{j=1}^n \mathbb{1}(Z_j > Z_i)(\nabla_{\boldsymbol{\beta}} S(\boldsymbol{X}_i;\boldsymbol{\beta}) - \nabla_{\boldsymbol{\beta}} S(\boldsymbol{X}_j;\boldsymbol{\beta}))K^{(1)}\left(\frac{S(\boldsymbol{X}_i;\boldsymbol{\beta}) - S(\boldsymbol{X}_j;\boldsymbol{\beta})}{h_n}\right)}{\frac{1}{n}\sum_{j=1}^n \mathbb{1}(Z_j > Z_i)}.$$

We next prove that the proposed adaptive Lasso penalty estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent, and present the selection consistency and asymptotic normality results. Let $\mathcal{I}_0 = \{k : \beta_k \neq 0, k = 1, \ldots, p\}$ be the set of important covariates. We denote $\hat{\mathcal{I}} = \{k : \hat{\beta}_k \neq 0, k = 1, \ldots, p\}$ as the set of selected covariates. Note that $k = 1$ is always included in both sets since we fix $\beta_1 = 1$. Without loss of generality, we write $\boldsymbol{\beta_0} = (\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{1n}^T, \hat{\boldsymbol{\beta}}_{2n}^T)^T$.

THEOREM 2: *Assume that Conditions (C1) and (C2) hold, If* $\sqrt{n}\lambda_n = O_p(1)$ *as* $n \to \infty$*, then* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\|_2 = O_p(n^{-1/2})$*. If* $\sqrt{n}\lambda_n \to 0$ *and* $n\lambda_n \to \infty$*, then the adaptive Lasso estimator* $\hat{\boldsymbol{\beta}}$ *satisfies* $\hat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$ *and* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \to N(\mathbf{0}, \boldsymbol{V_2})$*, in which* $\boldsymbol{V_2} = \boldsymbol{\Sigma_{11}^{-1}\Sigma_{21}\Sigma_{11}^{-1}}$*, and* $\boldsymbol{\Sigma_{11}}$ *and* $\boldsymbol{\Sigma_{21}}$ *are the leading* $d \times d$ *submatrices of* $\boldsymbol{\Sigma_1}$ *and* $\boldsymbol{\Sigma_2}$*.*

The above theorem asserts that if λ_n is chosen appropriately, then the penalization estimator enjoys the oracle property in the sense that it performs as well as the maximum likelihood estimator under the correct model (Donoho and Johnstone, 1994). In practice, λ_n can be chosen based on cross-validation; see more details in Section 5. The asymptotic results in Theorem 1 hold under mild conditions on the true score system, including linear and non-linear models. See the discussion and the proofs of Theorems 1 and 2 in the Supplementary File.

5. Simulation

5.1 *Score system without variable selection*

We first considered low-dimensional cases without variable selection. The data were generated under the following five scenarios:

Case 1. For each patient, we generated two independent biomarkers: X_1 from a uniform distribution on $(0, 1)$ and X_2 from a standard normal distribution. The survival time was generated from a mixture of two Weibull distributions $Z\text{Weibull}(2, 1) + (1 - Z)\text{Weibull}(1, 2)$, where Z followed a Bernoulli distribution with a success probability of $0.9\mathbb{1}(X_1 + 3X_2 \leqslant .5)$.

Case 2. In addition to the two biomarkers generated from Case 1, we generated another three redundant independent biomarkers: X_3 from a beta distribution $\text{Beta}(2, 5)$, X_4 from a Bernoulli distribution, Bernoulli(.3), and X_5 from a uniform distribution on $(-1, 1)$. The survival time was generated in exactly the same way as in Case 1.

Case 3. We generated X_1 from a uniform distribution on $(.8, 1.2)$ and an independent random preventive intervention assignment A following a Bernoulli distribution Bernoulli(.5). We also included the interaction of A and X_1 in the score system. The survival time was generated from a Weibull distribution with the following parameters:

$$\theta_{\text{shape}} = (1 - A)(2 - X_1) + A(1 + X_1), \quad \theta_{\text{scale}} = (1 - A)(3 - X_1) + AX_1.$$

Case 4. We considered the same setting as in Case 1 except that the success probability of Z was changed into $0.9\mathbb{1}(X_1^2 + 1.5X_2 \leqslant 1/3)$.

Case 5. We generated X_1 from $\text{Uniform}(.5, 1.5)$ and an independent A from Bernoulli(.5). The survival time T was generated by $\log T \sim N(-AX_1 - (1 - A)(3.5 - 1.5X_1), .5^2)$.

For each of 1000 simulated data sets, we used sample sizes of $n = 400$ and 1000, and randomly partitioned 75% of the patients into a training set and 25% into a test set. An independent censoring time was generated from a uniform distribution on $(0, \tau^*)$ such that either 20% or 40% of the patients were censored. The value of τ^* was determined numerically such that the empirical censoring rate was close to the desired percentage, such as 20%. For the training data, we used the exact method, the kernel method, the Cox model, the boosting method that maximizes the concordance probability (Mayr and Schmid, 2014),

the time-varying logistic regression model using inverse probability weighting (IPW) (Zheng et al., 2006), and a naive logistic model that discards the censored observations, to derive five risk score formulae. In the kernel method, we chose a fixed bandwidth $h_n = n^{-1/3}$, based on the recommendation by Jones (1990) for the optimal choice of bandwidth for density estimation. We applied risk score formulae to both the training and test data. Based on the resulting risk scores, we equally divided the patients into two groups (high-risk vs low-risk), and calculated the mean survival time for each group. The performances were evaluated by the differences in the mean survival times between the high-risk and low-risk groups, which were denoted by D^{test}_{mean}. We also calculate the IAUC using the method in Shen et al. (2015).

We summarize the results in Table 1. For all cases, the values of D_{mean} and IAUC obtained from the kernel smoothing method were very close to those obtained from the exact method, suggesting that the kernel approximation worked well for the non-smooth object function. The proposed methods (kernel and exact), the boosting method and the IPW logistic method all outperform the Cox model in most scenarios. This is expected because (1) the proportional hazard assumption is violated in these cases; and (2) the proposed method and the boosting method maximize the AUC to separate high-/low-risk groups as much as possible, while the Cox model maximizes the partial likelihood, which targets a different problem. The naive logistic method does not perform well for small sample sizes and high censoring rate (40%). This is expected since it does not use the censored observations.

For each case, we also calculated the IAUC using the underlying true score system. It can be seen that both the boosting method and the proposed method produced IAUC values close to the oracle truth, which confirms their prediction accuracy. The IPW logistic method does not work well enough when the proportional odds assumption is strongly violated, as in Cases 1–3, and when censoring is high (40%). For Case 4, by comparing the results from the true model (nonlinear) and the mis-specified model (linear), we notice that the proposed

method is quite robust to violations of model assumptions. Compared to the other methods, our method has good performances in all cases in terms of discriminating high-/low-risk groups. This is because the proposed method fully utilizes the covariate information from the censored subjects. We also considered the sensitivity analysis for the bandwidth; the results suggested that our method was not sensitive to the choice of bandwidth within a reasonable range.

[Table 1 about here.]

5.2 *Variable selection examples*

We next evaluated the performance of the penalized variable selection method. The data were generated from the following four situations.

Case 6 (weak correlation). We generated a biomarker X_1 from Uniform$(0, 1)$ and independently generated ten biomarkers $X_2, \ldots, X_{11}$ from a joint normal distribution with mean 0, variance 1 and correlation $\text{corr}(X_i, X_j) = 0.2\mathbb{1}(i \neq j)$. We also generated 5 i.i.d. biomarkers $X_{12}, \ldots, X_{16}$ from an exponential distribution with mean one. The survival time T was generated from a log-normal distribution $\log T \sim N(5 - 3X_1 - X_2, 0.04)$.

Case 7 (moderate correlation). We generated 20 biomarkers $X_1, \ldots, X_{20}$ from a joint normal distribution with mean 0, variance 1 and correlation $\text{corr}(X_i, X_j) = 0.5^{|i-j|}$. The survival time followed $\log T \sim N(2 - X_1 - 0.5X_2 - 0.5X_3, 0.04)$.

Case 8 (treatment interaction). We generated 15 biomarkers $X_1, \ldots, X_{15}$ from a joint normal distribution with mean 0, variance 1 and correlation $\text{corr}(X_i, X_j) = 0.15\mathbb{1}(i \neq j)$. An independent treatment indicator A was generated from a Bernoulli(0.5) distribution. The survival time followed $\log T \sim N(2.5 - X_1 - 0.5X_2 - 0.5X_1A - X_3A, 0.04)$.

Case 9 (strong correlation and treatment interaction) We generated $X_1, \ldots, X_{10}$ from a joint normal distribution with mean 0, variance 1 and correlation $\text{corr}(X_i, X_j) = 0.8^{|i-j|}$. We also

considered an independent indicator A from a Bernoulli(0.5) distribution. The survival time was generated in the same way as in Case 8.

For each case, we considered 1000 replications and summarized the selection frequency of each individual biomarker and the proportion of exactly selecting the correct model in Table 2. We listed the estimated coefficients for selected biomarkers. We randomly partitioned the data into training (40%), validation (40%) and test (20%) sets. We let the tuning parameter λ take values in $\{5, 7, 10, 15, 20, 50, 70, 100\}$ and chose the optimal one by maximizing the median survival difference (D_{med}) on the validation set.

As shown in Table 2, the proposed penalization method showed a good selection performance, particularly when the sample size was large. For example, under Case 8 with sample size of 1000, the selection frequencies for true variables X_2, X_1A and X_3A obtained from the proposed method were greater than 96%, and the true model was selected more than 80% of the time. The selection frequencies for false variables A and X_2A were very low (at most 7%). The method also provided accurate estimation of the coefficients. Taking Case 7 as an example, the estimation biases were at most 0.20 for $n = 500$, and 0.12 for $n = 1000$. Case 9 was a challenging situation because of the high correlation and the treatment interaction. Therefore the selection frequency of the unimportant variable X_2A was high since X_1A, X_3A and X_2 were in the true model and are highly correlated. Still, the selection frequency and estimated coefficients of all other variables were satisfactory. To evaluate the prediction performance of our method, we compared the difference in the mean and median survival times for the high-risk and low-risk groups as determined by the proposed penalized method and by the oracle model, in which we fitted the model using the true set of biomarkers. The results on the test and training data (reported in brackets) are summarized in Table 3 based on 1000 replications. The prediction performance after variable selection seemed satisfactory

in the sense that the survival time difference was very close (Cases 7–9 in particular) to that obtained from the oracle model for both low (20%) and high (40%) censoring rates.

Note that dividing the patients equally into two groups and looking at their survival time difference is a convenient way to evaluate the discrimination ability of the score systems, but is not the only way to determine patient subgroups. In application, a possible alternative is to apply cluster analysis for the obtained scores to determine the cut-offs to use in dividing patients into subgroups.

[Table 2 about here.]

[Table 3 about here.]

6. Data application

In the Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) trial, a total of 1050 patients with chronic hepatitis C were enrolled and randomized to receive half-dose pegylated interferon or no treatment, and were followed up for a median of 6.1 (maximum 8.7) years to monitor the development of HCC and liver failure. At entry, clinicians obtained measurements from each patient, including blood cell count, liver panel (e.g., albumin, aspartate aminotransferase, alanine aminotransferase and bilirubin), and AFP/DCP levels. Time to HCC was one of the planned trial outcomes; hence, the HALT-C trial provided an opportunity to identify risk factors associated with the development of HCC in a U.S. cohort with chronic hepatitis C and advanced liver fibrosis or cirrhosis. There were 88 patients who met the criteria for HCC. We aimed to identify risk factors associated with the development of HCC and to construct a score formula for risk prediction.

We applied the method in Section 3 to the HALT-C data. In addition to the aforementioned variables, we included demographic information (age [mean=50.1], gender [female 29%], race [white 71.6%, Hispanic 18.2%, black and other 10.2%]), body mass index (BMI), smoking

status, alcohol consumption, baseline Child-Turcotte-Pugh (CTP) score (measures cirrhosis complications and synthetic function of the liver), baseline liver biopsy Ishak score (measures the stage of fibrosis on a scale from 0 to 6) and their interactions in our analysis. We treated the time from the initial entry to the development of HCC as the primary outcome.

We first evaluated the risk prediction abilities of individual factors and combined a subset of factors without variable selection. To perform the assessment, we considered five-fold cross-validation by randomly partitioning the data set into five equal folds, choosing four for model fitting and using the other one to calculate the IAUC for the 1-year risk prediction with the method of Shen et al. (2015). This procedure is repeated five times for each fold. The average of the IAUCs together with their associated standard errors are presented in Table 4. Among all factors, AFP had the highest IAUC (.581), confirming the clinical experience of treating AFP as one of the most useful biomarkers for diagnosis of HCC as a risk factor that predicts HCC development. In contrast, bilirubin had IAUC values close to the non-informative value of .50, suggesting that it did not have good prediction performance when evaluated individually. The risk score obtained by combining risk factors may lead to a significant increment in the IAUC. For example, the IAUC value increased from .581 to .836 by combining the Ishak fibrosis score with AFP. Similarly, by combining DCP with race, the IAUC value increased from .530 to .726.

We then used the penalization method in Section 3.3 to identify the important risk factors and build a score system. To reduce the potential bias associated with overfitting, we used five-fold cross-validation to choose the tuning parameter λ in a pre-specified set that had the highest estimated mean survival time difference between the low-risk and high-risk groups. For simplicity, we fixed the bandwidth of the normal kernel at $n^{-1/3}$. The identified set of significant factors in the score formula included the main effects of AFP, DCP, albumin, platelets, ALK, age, gender, BMI, Ishak score, and the interaction effect of AFP * gender,

AFP * Race and DCP * Race. The IAUC by the score system was .943 with a standard error of .015, which increased the IAUC of the single AFP by 62%.

For comparison, we also considered a risk score that consisted of the set of variables provided by Lok et al. (2009) using Cox model, and a risk score without the interaction DCP * Race. This showed that the interaction played a role in the score system as the IAUC dropped from .943 to .856 when the interaction was removed. The proposed set of variables increased the IAUC for the model of Lok et al. (2009) by 29%.

We then evaluated the 5-year risk prediction performance. The estimated IAUC of the score system was .720 with a standard error of .045. Comparing to the 1-year risk prediction, the value of IAUC was decreased by 24%. This was expected because we were only using the baseline information in the score system. To better illustrate, in Figure 1, on the left side we plotted the estimated AUC curve of the score system with 95% confidence bands and that of the single biomarker AFP for the entire study period. This plot shows that the score system had a significant advantage in estimation accuracy compared to AFP, as the AUC of the proposed score was always above that of AFP. In the first two years, the estimated AUC of the score system was above .8, indicating a good prediction performance. However, the predictive ability of that score decreased quickly over time, implying that the risk score was not that informative for long-term risk prediction, and that longitudinally measured values of the risk factors should be utilized to update and improve the risk assessment. We further divided the patients equally into high-risk and low-risk groups based on their scores and plotted their corresponding Kaplan-Meier survival curves; the proposed method managed to separate these two subgroups well.

[Figure 1 about here.]

[Table 4 about here.]

7. Discussion

In this paper, we proposed new methods for selecting important risk factors and constructing a clinically useful score system. The proposed method builds on the idea of maximizing a global average of the time-dependent AUC. Hence, it avoids the need of making strong model assumptions on the time-to-event outcomes while boosts the overall prediction performance.

So far we have assumed that no two patients have the same failure event time in the data. In the presence of ties, the proposed method can be extended by considering all possible orderings (or randomly choosing one possible ordering), adding a small amount to the failure times and then summing up the likelihoods under different orderings.

In addition to ROC-based approaches, logistic regression provides another valuable alternative to score system estimation (Vexler et al., 2016). To fully utilize the censored observations and time-to-event outcomes, inverse probability weighting (IPW) and Bayesian models can be used. In our simulations, we have implemented a naive logistic regression approach and a time varying logistic regression method with IPW for score system estimation. It can be seen that the results from the logistic models are comparable to those from our approach when implemented appropriately. It is of future interest to develop a dynamic Bayesian logistic regression approach for score system estimation and covariate selection.

To improve the precision of risk assessments, molecular/genetic data have been included in the risk profiles for cancer research. However, measuring such biomarkers can be expensive and labor intensive; hence, the collection of biomarker information for each subject may be infeasible. It will be interesting to adapt the current framework to more complicated designs such as nested case-control and stratified case-cohort designs (Cai and Zheng, 2013). In this paper, we assume that the score formula $S(\boldsymbol{X}, \boldsymbol{\beta})$ follow a known functional form. In future work, it will be of interest to develop methods to test the validity of these model assumptions. Another important area for future research is nonparametric estimation of functions ψ in

(3) by considering basis expansion techniques such as splines and wavelets. For biomarker selection, we adopt a penalization method. It is also possible to consider Bayesian approaches with variable selection priors (e.g., spike-and-slab and horseshoe).

8. Supplementary Materials

Additional numerical results and technical proofs referenced in Sections 3.2 and 4, together with the simulation code and data are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgments

The authors thank the editor, the associate editor and two reviewers for their constructive comments that have greatly improved the initial version of this paper. This work was supported in part by a Cancer Center Support Grant from the National Institutes of Health (CA016672), a grant from the Cancer Prevention Research Institute of Texas (RP150587) and UCI CORCL award.

References

Cai, T. and Zheng, Y. (2013). Resampling procedures for making inference under nested case-control studies. *Journal of the American Statistical Association* **108,** 1532–1544.

Chang, Y. C. (2013). Maximizing an roc-type measure via linear combination of markers when the gold reference is continuous. *Stat. Med.* **32,** 1893–1903.

Chen, B., Li, P., Qin, J., and Yu, T. (2016). Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association* **111,** 861–874.

Chen et al. (2015). Empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers. *Computational Statistics and Data Analysis* **82,** 186–198.

Cui, R., et al. (2003). Diagnostic value of protein induced by vitamin k absence (pivkaii) and

hepatoma-specific band of serum gamma-glutamyl transferase (ggtii) as hepatocellular carcinoma markers complementary to alpha-fetoprotein. *Br J Cancer.* **88,** 1878–1882.

D'Avanzo, A., et al. (2004). Prognostic scoring systems in patients with follicular thyroid cancer: a comparison of different staging systems in predicting the patient outcome. *Thyroid* **14,** 453–458.

Davis, G. L., et al. (2010). Aging of hepatitis c virus (hcv)-infected persons in the united states: a multiple cohort model of hcv prevalence and disease progression. *Gastroenterology* **138,** 513–521.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81,** 425–455.

Etzioni, R., Kooperberg, C., Pepe, M. S., Smith, R., and Gann, P. H. (2003). Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics* **4,** 523–538.

Everhart, J. E. and Ruhl, C. E. (2009). Burden of digestive diseases in the united states part iii: Liver, biliary tract, and pancreas. *Gastroenterology* **136,** 1134–1144.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1,** 302–332.

Greenberg, P., Cox, C., LeBeau, M. M., and et. al. (1997). International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood* **89,** 2079–2088.

Greenberg, P. L., Tuechler, H., Schanz, J., and et. al. (2012). Revised international prognostic scoring system for myelodysplastic syndromes. *Blood* **120,** 2454–2465.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics* **61,** 92–105.

Jones, M. C. (1990). The performance of kernel density functions in kernel distribution

function estimation. *Statistics & Probability Letters* **9,** 129–132.

Kadalayil, L., et al. (2013). A simple prognostic scoring system for patients receiving transarterial embolisation for hepatocellular cancer. *Annals of Oncology* **24,** 2565–2570.

Liu, A., Schisterman, E. F., and Zhu, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine* **24,** 37–47.

Lok et al. (2009). Incidence of hepatocellular carcinoma and associated risk factors in hepatitis c-related advanced liver disease. *Gastroenterology* **136,** 138–148.

Lok et al. (2010). Des-gamma-carboxy prothrombin and alpha-fetoprotein as biomarkers for the early detection of hepatocellular carcinoma. *Gastroenterology* **138,** 493–502.

Ma, S. and Huang, J. (2005). Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21,** 4356–4362.

Mayr, A. and Schmid, M. (2014). Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations. *PLoS One.* **9,** e84483.

McIntosh, M. and Pepe, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics* **58,** 657–664.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7,** 308–313.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62,** 221–229.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1,** 123–140.

Santi, V., et al. (2010). Semiannual surveillance is superior to annual surveillance for the detection of early hepatocellular carcinoma and patient survival. *Journal of hepatology* **53,** 291–297.

Sanyal, A. J., et al. (2006). The prevalence and risk factors associated with esophageal varices

in subjects with hepatitis c and advanced fibrosis. *Gastrointest Endosc.* **64,** 855–864.

Shen, W., Ning, J., and Yuan, Y. (2015). A direct method to evaluate the time-dependent predictive accuracy for biomarkers. *Biometrics* **71,** 439–449.

Sidransky, D. (2002). Emerging molecular markers of cancer. *Nat. Rev. Cancer* **2,** 210–219.

Song, P., Tang, W., and Kokudo, N. (2014). Serum biomarkers for early diagnosis of hepatocellular carcinoma. *Translational Gastrointestinal Cancer* **3,** 103–105.

Song, P., et al. (2013). Biomarkers: Evaluation of screening for and early diagnosis of hepatocellular carcinoma in japan and china. *Liver Cancer.* **2,** 31–39.

Su, J. and Liu, J. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association.* **88,** 1350–1355.

Vexler, A., Hutson, A. D., and Chen, X. (2016). *Statistical Testing Strategies in the Health Sciences*. Chapman and Hall/CRC.

Yuan, Z. and Ghosh, D. (2008). Combining multiple biomarker models in logistic regression. *Biometrics* **64,** 431–439.

Zeng, D. and Lin, D. Y. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association* **102,** 1387–1396.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94,** 691–703.

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.

Zhu, W. W., et al. (2013). Evaluation of midkine as a diagnostic serum biomarker in hepatocellular carcinoma. *Clin Cancer Res.* **19,** 3944–3954.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

Figure 1. HALT-C data analysis.

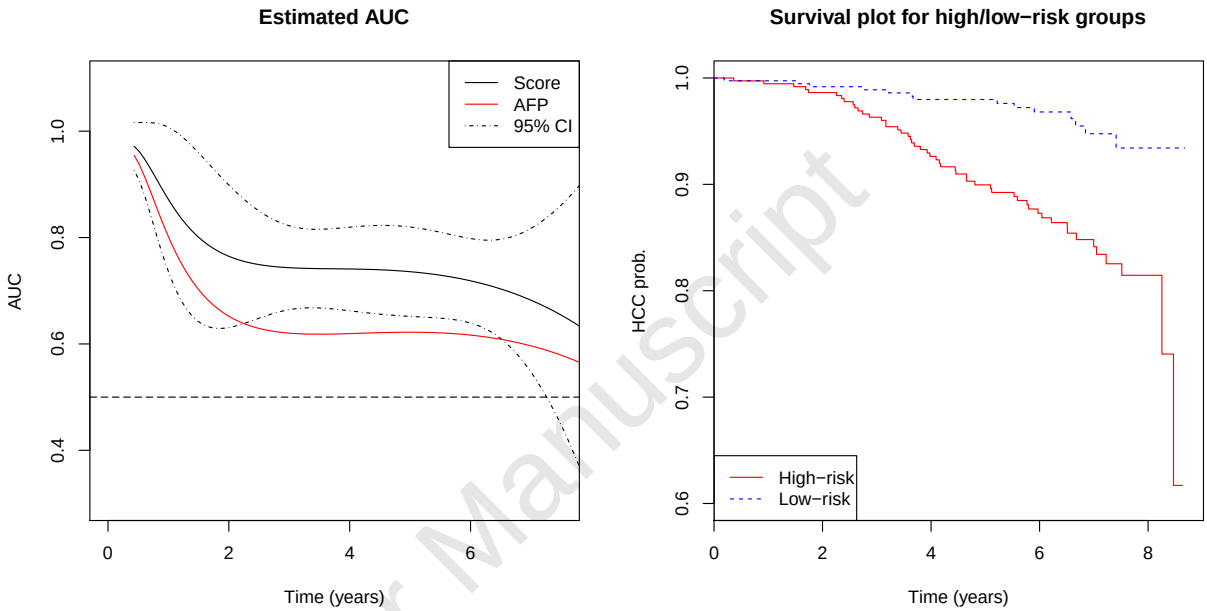

Table 1

Simulation results for Cases 1–5: Mean survival time difference between low-risk and high-risk groups and the estimated IAUC, obtained from the kernel-smoothing method (Ker), the exact method (Ext), the Cox model (C), the boosting method (M), the time-varying logistic regression model (Z), and the naive logistic regression method (L) based upon 1000 replications

Setting			D^{test}_{mean}						IAUC						
case	n	cen	Ker	Ext	C	M	Z	L	True	Ker	Ext	C	M	Z	L
1	200	.2	.42	.42	.19	.32	.18	.07	.61	.58	.58	.53	.59	.53	.51
		.4	.24	.27	.10	.18	.05	.04	.59	.57	.56	.52	.59	.51	.51
	400	.2	.42	.41	.22	.35	.21	.07	.60	.59	.59	.53	.60	.53	.51
		.4	.27	.27	.10	.20	.07	.02	.59	.57	.57	.52	.59	.52	.51
	1000	.2	.56	.60	.30	.37	.37	.09	.61	.60	.60	.55	.60	.55	.51
		.4	.39	.40	.17	.21	.11	.05	.59	.59	.59	.54	.59	.52	.51
2	200	.2	.28	.34	.12	.25	.16	.03	.60	.58	.56	.52	.58	.53	.50
		.4	.16	.20	.06	.14	.04	.00	.59	.58	.55	.52	.58	.52	.50
	400	.2	.40	.37	.15	.33	.21	.04	.60	.59	.58	.53	.60	.54	.51
		.4	.25	.25	.07	.18	.05	.03	.59	.58	.57	.53	.58	.51	.50
	1000	.2	.52	.51	.26	.38	.33	.04	.60	.60	.60	.54	.60	.55	.51
		.4	.33	.34	.15	.21	.05	.02	.59	.58	.58	.53	.59	.51	.51
3**	200	.2	.87	.87	.28	.66	.19	.23	.62	.64	.62	.54	.63	.53	.54
		.4	.55	.55	.27	.39	.20	.00	.59	.63	.60	.57	.61	.55	.50
	400	.2	.89	.88	.40	.70	.36	.34	.62	.64	.62	.56	.63	.58	.56
		.4	.57	.57	.36	.42	.23	.02	.59	.63	.60	.58	.61	.54	.51
	1000	.2	.94	.93	.60	.72	.48	.45	.62	.64	.62	.59	.63	.61	.58
		.4	.61	.60	.51	.43	.27	.06	.59	.63	.60	.61	.61	.54	.52
4	200	.2	.43	.46	.33	.30	.33	.11	.61	.58	.58	.55	.59	.55	.52
		.2*	.35	.46	.33	.29	.35	.10	.61	.57	.58	.55	.59	.55	.52
		.4	.29	.27	.18	.15	.11	.09	.59	.57	.57	.54	.58	.52	.52
		.4*	.22	.27	.17	.16	.11	.09	.59	.56	.57	.54	.58	.53	.51
	400	.2	.48	.54	.38	.34	.41	.11	.60	.59	.59	.56	.60	.56	.52
		.2*	.35	.54	.40	.34	.42	.11	.60	.57	.59	.56	.60	.56	.52
		.4	.30	.31	.20	.26	.13	.08	.59	.58	.58	.54	.59	.53	.51
		.4*	.23	.32	.23	.19	.15	.07	.59	.56	.58	.55	.59	.53	.51
	1000	.2	.62	.66	.57	.55	.58	.17	.60	.60	.60	.58	.60	.59	.52
		.2∗	.42	.66	.58	.36	.59	.15	.60	.58	.60	.58	.60	.59	.52
		.4	.38	.40	.30	.29	.20	.07	.59	.59	.59	.56	.59	.54	.51
		.4∗	.29	.42	.33	.20	.20	.08	.59	.57	.59	.56	.59	.54	.52
5	200	.2	.23	.21	.08	.18	.20	.10	.75	.68	.66	.56	.60	.61	.55
		.4	.21	.18	.04	.15	.08	.04	.71	.67	.66	.51	.55	.56	.51
	400	.2	.24	.24	.12	.18	.22	.14	.75	.68	.68	.56	.59	.64	.56
		.4	.21	.21	.06	.15	.11	.05	.71	.67	.67	.50	.55	.59	.51
	1000	.2	.25	.25	.19	.19	.23	.20	.75	.68	.68	.58	.59	.66	.57
	1000	.4	.22	.22	.09	.16	.13	.12	.71	.67	.667	.52	.55	.60	.53

D^{test}_{mean}: mean difference in the test data; IAUC: integrated area under the curve.

*: Case 4, the true model contains a nonlinear term. We fit both the linear model (results denoted by $*$ in the table) and the nonlinear model to evaluate the robustness of the proposed method under model assumption violations.

**: Case 3, the true score system is two-dimensional, hence there is no "true" one-dimensional score. Therefore we use the interaction $X_1(2A-1)$ to generate the "True" value for IAUC.

Table 2
Simulation results for Cases 6–9: Selection frequency and estimated coefficients for Cases 6–9 based upon 1000 replications.

Case 6		Selection frequency (%)						Estimated coefficients		
n	cen(%)	X_2	X_3	X_4	X_{12}	X_{13}	True Model	X_2	X_3	X_{14}
500	20	99.8	1.5	2.5	.9	2.2	79.4	.380	.015	.008
	40	99.9	3.9	3.8	2.4	2.2	66.4	.377	.010	.000
1000	20	99.6	0.7	0.8	0.3	0.3	91.8	.363	-.002	.003
	40	99.6	2.5	2.2	0.9	1.1	80.0	.361	-.002	.000

Case 7		Selection frequency (%)						Estimated coefficients		
n	cen%	X_2	X_3	X_4	X_5	X_{10}	True Model	X_2	X_3	X_4
500	20	100	100	6.8	6.9	4.6	55.7	.520	.511	.002
	40	100	100	9.5	8.4	8.0	52.1	.521	.511	.003
1000	20	100	100	0.5	0.2	0.5	91.6	.511	.504	.010
	40	100	100	1.9	1.9	1.1	82.2	.512	.505	.009

Case 8		Selection frequency (%)						Estimated coefficients		
n	cen%	X_2	A	X_1A	X_2A	X_3A	True Model	X_2	X_1A	X_3A
500	20	99.6	10.8	99.8	9.5	100	46.6	.518	.566	1.044
	40	99.8	9.4	96.0	18.0	100	37.9	.515	.586	1.048
1000	20	99.6	0.6	99.9	1.5	100	93.2	.510	.535	1.022
	40	99.8	1.2	96.8	6.8	100	80.8	.508	.548	1.024

Case 9		Selection frequency (%)						Estimated coefficients		
n	cen%	X_2	A	X_1A	X_2A	X_3A	True Model	X_2	X_1A	X_3A
500	20	100	6.9	100	62.8	100	16.3	.592	.612	1.076
	40	99.9	9.5	100	53.9	100	18.0	.597	.638	1.089
1000	20	100	0.9	100	51.8	100	32.7	.551	.569	1.044
	40	100	1.4	100	45.7	100	28.5	.552	.580	1.056

Table 3

Simulation results for Cases 6–9: Mean/median survival time differences using the proposed penalized method (penal) compared with the oracle model (oracle) based upon 1000 replications

			$D^{test}_{\text{mean}}(D^{train}_{\text{mean}})$		$D^{test}_{\text{med}}(D^{train}_{\text{med}})$	
Case	n	cen(%)	penal	oracle	penal	oracle
6	500	20	69.4 (70.2)	96.7 (97.0)	53.5 (53.5)	71.0 (70.5)
		40	43.7 (43.9)	67.3 (67.4)	50.1 (52.9)	64.9 (68.8)
	1000	20	71.4 (71.6)	98.5 (98.7)	54.6 (54.2)	71.3 (70.5)
		40	45.5 (45.3)	67.5 (67.3)	53.6 (54.0)	69.2 (70.2)
7	500	20	24.7 (25.0)	27.7 (28.0)	17.5 (17.6)	19.1 (19.3)
		40	13.9 (13.9)	16.9 (16.9)	15.0 (16.4)	16.5 (18.0)
	1000	20	25.3 (25.5)	28.0 (27.9)	17.8 (17.7)	19.2 (19.0)
		40	14.2 (14.1)	16.7 (16.6)	16.7 (17.3)	17.7 (18.4)
8	500	20	40.3 (40.5)	46.2 (46.0)	27.4 (27.4)	31.0 (30.4)
		40	22.3 (22.3)	26.8 (26.9)	24.2 (26.0)	26.2 (28.3)
	1000	20	40.8 (41.1)	47.5 (47.2)	27.3 (27.5)	31.0 (30.4)
		40	22.1 (22.2)	26.4 (26.4)	25.7 (26.9)	25.8 (28.1)
9	500	20	66.5 (65.5)	69.1 (69.6)	41.4 (40.0)	42.0 (42.2)
		40	30.7 (30.5)	32.4 (32.5)	31.4 (34.6)	31.5 (34.8)
	1000	20	66.9 (66.9)	69.6 (69.3)	40.8 (40.5)	42.3 (42.0)
		40	30.8 (30.6)	32.7 (32.7)	34.7 (37.0)	35.0 (37.6)

D^{test}_{mean}: mean difference in the test data; D^{train}_{mean}: mean difference in the training data; D^{test}_{med}: median difference in the test data; D^{train}_{med}: median difference in the training data.

Table 4
One-year risk prediction for HALT-C data: IAUC and associated standard error (SE) for different set of risk factors

Factors	IAUC	SE
AFP	.581	.037
ALK	577	.037
AST	.561	.040
DCP	.530	.039
Bilirubin	.503	.035
AFP, Ishak	.836	.029
DCP, Race	.726	.031
Score system including AFP, DCP, ALK, Ishak, Age, BMI, Gender, Drink, Platelets, AST, Albumin, CTP, AFP*Gender, AFP*Race and DCP*Race	.943	.015
Score system without the interaction term DCP*Race	.856	.015
Lok et al. (2009)'s formula (Age, Race, Smoke, Platelets, ALK)	.731	.016