Published in final edited form as:

Biometrics. 2016 September ; 72(3): 897–906. doi:10.1111/biom.12470.

Addressing issues associated with evaluating prediction models for survival endpoints based on the concordance statistic

Ming Wang and

Department of Public Health Sciences, College of Medicine, Pennsylvania State University, Hershey, PA, U.S.A

Qi Long

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, U.S.A

Ming Wang: mwang@phs.psu.edu; Qi Long: qlong@emory.edu

Summary

Prediction models for disease risk and prognosis play an important role in biomedical research, and evaluating their predictive accuracy in the presence of censored data is of substantial interest. The standard concordance (c) statistic has been extended to provide a summary measure of predictive accuracy for survival models. Motivated by a prostate cancer study, we address several issues associated with evaluating survival prediction models based on c -statistic with a focus on estimators using the technique of inverse probability of censoring weighting (IPCW). Compared to the existing work, we provide complete results on the asymptotic properties of the IPCW estimators under the assumption of coarsening at random (CAR), and propose a sensitivity analysis under the mechanism of noncoarsening at random (NCAR). In addition, we extend the IPCW approach as well as the sensitivity analysis to high-dimensional settings. The predictive accuracy of prediction models for cancer recurrence after prostatectomy are assessed by applying the proposed approaches. We find that the estimated predictive accuracy for the models in consideration is sensitive to NCAR assumption, and thus identify the best predictive model. Finally, we further evaluate the performance of the proposed methods in both settings of low-dimensional and high-dimensional data under CAR and NCAR through simulations.

Keywords

Predictive accuracy; Survival endpoint; Concordance statistic; Inverse probability of censoring weighting; Coarsening at random; Noncoarsening at random; Sensitivity analysis

Supplementary Materials

Web Appendices A, B and C referenced in Sections 2 and 4 are available with this paper at the *Biometrics* website on Wiley Online Library.

The content is solely the responsibility of the authors and does not represent the views of the NIH.

1. Introduction

Prediction models play an important role for assessing disease risk and prognosis (Freedman et al., 2005). They can be used to identify individuals at high risk, plan interventional trials, and improve personalized prevention and treatment strategies. Furthermore, they can also estimate the population burden, the cost of disease, and the impact of potential interventions or treatments. As a result, research in prediction models for disease risk and prognosis has drawn substantial interest in recent years (Freedman et al., 2009; Optenberg et al., 1997).

1.1 A Prostate Cancer Study

This work is motivated by a prostate cancer study where the outcome of interest is the time (months) to prostate cancer recurrence after prostatectomy. One of the main objectives is to find out an effective prediction model of future tumor recurrence after surgery to determine whether an immediate adjuvant therapy is warranted, thus a valid predictive accuracy evaluation is of the utmost importance. In this study, patients diagnosed with prostate cancer underwent radical prostatectomy, and their prostatectomy specimens were immediately collected after the surgery to be formalin-fixed and paraffin-embedded (FFPE). Subsequently, the RNA samples were isolated from these specimens with known long-term outcomes, and then mRNA and microRNA expression profiling were performed to obtain gene expression data by using a custom-designed panel of 1,536 probes (called biomarkers thereafter) of 522 prostate cancer relevant genes. In particular, a novel expression profiling platform, the DASL (cDNA-mediated Annealing, Selection, extension and Ligation) assay, was applied to allow a quantitative analysis of RNA from FFPE samples. In addition, the relevant clinical biomarkers were also collected, among which prostate-specific antigen (PSA) and Gleason scores were taken as key consideration due to their known association with prostate cancer risk. Using an earlier version of the prostate cancer data, Long et al. (2011) developed a robust partial linear accelerated failure time model relating recurrence risk to genomic biomarkers and the above two clinical risk factors. Later on, Long et al. (2011) identified a panel of protein-coding and microRNA biomarker signatures to predict cancer recurrence following prostatectomy. However, in these work, the predictive accuracy of the models was evaluated under a strong, likely an unrealistic assumption of independent censoring.

Since censoring is a special case of coarsened data, throughout we adopt the terminology from coarsened data to define the censoring mechanism (Heitjan and Rubin, 1991): 1) coarsening complete at random (CCAR), i.e., censoring is independent of the failure times; 2) coarsening at random (CAR), i.e., censoring is independent of the failure times conditional on the observed covariates; and 3) non-coarsening at random (NCAR), i.e., censoring is dependent on the failure times conditional on the observed covariates. The goal of our work is to evaluate predictive accuracy of the models adjusted for the censoring mechanisms of CAR and NCAR in the settings of low-dimensional and high-dimensional data.

1.2 Assessing Predictive Accuracy

In order to identify the optimal prediction model for subsequent applications in practice, it is crucial to develop robust predictive accuracy metrics for assessing and comparing different prediction models (Steyerberg et al., 2010). While it is fairly straightforward to assess prediction performance when all data are fully observed, it tends to be considerably challenging in the presence of censored data, which are ubiquitous in biomedical and epidemiologic research. In addition, it is common to assess prediction models for the censored data in observational studies where the censoring mechanism is likely to be CAR or NCAR (Siannis et al., 2005). Even if assessment is performed using the data from clinical trials, the censoring mechanism could still be CAR or NCAR due to informative dropouts or missing (Scharfstein et al., 1999). For such situations, the prediction assessment can be biased due to the ignorance of adjustment on censoring mechanisms.

While a considerable amount of work has been reported on development of predictive accuracy metrics (D'Agostino et al., 1997), there has been limited work on predictive accuracy metrics for censored data. The standard concordance (c) statistic as one of predictive accuracy measures has been extended to censored data (Harrell et al., 1996). For instance, Pencina and D'Agostino (2004) provided some insight on the relationship between the c -statistic and Kendall's τ , and investigated its asymptotic normality; Gonen and Heller (2005) derived an analytical expression for the concordance probability under the framework of Cox Proportional Hazards (PH) models (Cox, 1972), which was robust to censoring but not invariant to monotone transformations of survival times. However, all of those have been developed under CCAR and/or limited to Cox PH models. As mentioned before, under CAR or NCAR, the standard c -statistic potentially leads to biased assessment of predictive accuracy or an estimate with a large variability (Uno et al., 2011; Gerds et al., 2013). Some recent work has attempted to modify the standard c -statistic. Uno et al. (2011) proposed a modified c -statistic by using the technique of inverse probability of censoring weighting (IPCW) that is valid under CCAR or CAR, where censoring weights were obtained by Kaplan-Meier (KM) estimates. Along similar lines, Gerds et al. (2013) proposed an alternative time-dependent IPCW c -statistic that is valid under CAR, which is closely related to our current work; though, they only provided the theoretical results on the consistency of the proposed estimator, and did not consider several complicated issues, such as NCAR and high-dimensional data.

As shown in the motivating study in Section 1.1, high-dimensional data such as genomic or proteomic data are commonly collected in biomedical research studies, presenting a challenge for evaluating prediction models because these data may provide additional information that is not captured by clinical variables. In particular, it is not trivial to incorporate the survival probabilities of censoring times particularly in the presence of high-dimensional data when estimating the IPCW c -statistic. While modern regularization methods have been developed including lasso (Tibshirani, 1997), ridge (Hoerl and Kennard, 1970), adaptive lasso (Zou, 2006), elastic net (Zou and Hastie, 2005) and so on, there is still not clear whether these techniques can improve the efficiency of the IPCW c -statistic. In the current work, we focus on lasso and elastic net, and figure out their applicability for obtaining the censoring weights when assessing prediction models in the high-dimensional

setting. Furthermore, considering that censoring is likely to be CAR or even NCAR in the motivating study, we make the following contributions in addressing several issues associated with evaluation of prediction models with survival endpoints based on the IPCW c -statistic. First, we provide more complete theoretical results on the asymptotic properties of the IPCW c -statistic under CAR. Second, we propose a sensitivity analysis approach under NCAR. Third, we extend the proposed approaches to the case of high-dimensional data.

The remainder of the article is organized as follows. In Section 2, we formulate the problem, and introduce several types of c -statistics for censored data. We provide the asymptotic results for the extended IPCW c -statistics under CAR, propose a sensitivity analysis under NCAR, and investigate extensions of the proposed approaches to the setting of high-dimensional data. In Section 3, we apply the proposed approaches to the motivating study on prostate cancer. In Section 4, simulation results for comparing different estimators are provided, and the advantages of the IPCW c -statistic are illustrated afterwards. Note that our simulation studies are more extensive compared to the existing work by considering both of Cox PH models and accelerated failure time (AFT) models as well as high-dimensional settings. The conclusion with a brief discussion is provided in Section 5.

2. Methodology

Let $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ denote a *training* data set and a *testing* data set, which are independent with each other. $\mathcal{D}_{train}$ containing n observations is used to build up a prediction model and $\mathcal{D}_{test}$ including m observations is utilized to evaluate its prediction performance, where m/n is bounded away from 0 and ∞ . $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are obtained through either randomly splitting one data set into two or using data sets collected in two independent studies. The censoring mechanisms in the former case are expected to be the same in both $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ whereas they may be different in the latter case. Without loss of generality, we assume that $m = n$, simplifying the derivation of the asymptotic results.

For observation i in $\mathcal{D} = (\mathcal{D}_{train}, \mathcal{D}_{test})$, we denote the true failure time by T_i which is subject to censoring, the censoring time by C_i , the observed survival time by $Y_i = \min(T_i, C_i)$, and a set of p covariates by $\mathbf{Z}_i$ that are potentially associated with T_i and/or C_i . We also define $\delta_i = I(T_i < C_i)$ as the event indicator. The observed data in $\mathcal{D}$ are $\{Y_i, \delta_i, \mathbf{Z}_i\}_{i=1}^{2n}$, where $(Y_i, \delta_i, \mathbf{Z}_i)$'s are assumed to be independent.

2.1 Predictive Accuracy Metric: Truncated c -statistic

Suppose that we are interested in assessing the predictive accuracy of any particular model for time-to-event data, denoted by $\mathcal{M}_1$, which is parameterized by β . $\mathcal{M}_1$ may be correctly-specified or mis-specified, a parametric or semi-parametric model, and a linear, partly linear, nonlinear, or varying coefficient regression model. For observation i , let S_i denote the predictive score under $\mathcal{M}_1$ with a higher value representing a higher survival probability. Thus, given any two independent observations 1 and 2, the predictive accuracy metric of interest for a pre-specified time point t is defined by the truncated c -statistic as

$$c(t) = Pr(S_1 < S_2 | T_1 < T_2, T_1 \leq t). \quad (1)$$

where $\alpha(t) = 1$ indicates the perfect discrimination. Of note, the truncated c -statistic is preferred for right censored data primarily because the tail part of the estimated survival function of the failure time is known to be unstable (Heagerty and Zheng, 2005; Uno et al., 2011; Gerds et al., 2013).

2.2 Estimation of Truncated c -statistic

The parameter estimates in $\mathcal{M}_1$, denoted by $\hat{\beta}$, are obtained using $\mathcal{D}_{train}$ through an estimation procedure, e.g., the partially likelihood approach for a Cox PH model. The corresponding estimated model is denoted by $\hat{\mathcal{M}}_1$. For observation i in $\mathcal{D}_{test}$, the estimated predictive score under $\mathcal{M}_1$ is defined as $\tilde{S}_i = \xi(\hat{\beta}; \mathbf{Z}_i)$, where $\xi(\cdot)$ is a smooth and monotone function determined by the model specification. For instance, if $\mathcal{M}_1$ is a Cox PH model, $h(t; \beta, \mathbf{Z}) = h_0(t) \exp(\beta^T \mathbf{Z})$ with $h_0(\cdot)$ as the baseline hazard function, we can define $\tilde{S}_i = -\hat{\beta}^T \mathbf{Z}_i$; if $\mathcal{M}_1$ is a parametric AFT model, $\log(t; \beta, \mathbf{Z}) = \beta^T \mathbf{Z} + \varepsilon$ with the random error ε following a given parametric distribution, we can define $\tilde{S}_i = \hat{\beta}^T \mathbf{Z}_i$. Assume that $\hat{\beta}$ converges in probability to a constant β_0 , even if $\mathcal{M}_1$ is incorrectly specified; this holds under a mild non-separable condition when $\mathcal{M}_1$ is a Cox PH model (Uno et al., 2011; Zeng et al., 2004). It follows that the true predictive score under $\mathcal{M}_1$ is $S_i = \xi(\beta_0; \mathbf{Z}_i)$.

Under CCAR, one estimator of truncated $\alpha(t)$ is defined as

$$\hat{c}_s^1(\hat{\beta}; t) = \frac{\sum_{i \neq j} \delta_i \delta_j I(\tilde{S}_i < \tilde{S}_j, Y_i < Y_j, Y_i \leq t)}{\sum_{i \neq j} \delta_i \delta_j I(Y_i < Y_j, Y_i \leq t)}, \quad (2)$$

where only the pairs with both observed survival times are used. If the pairs with only one censored outcome are also considered, it leads to the standard estimator of truncated $\alpha(t)$

$$\hat{c}_s^2(\hat{\beta}; t) = \frac{\sum_{i \neq j} \delta_i I(\tilde{S}_i < \tilde{S}_j, Y_i < Y_j, Y_i \leq t)}{\sum_{i \neq j} \delta_i I(Y_i < Y_j, Y_i \leq t)}. \quad (3)$$

Note that both of these two estimators are consistent under a set of restrictive conditions including CCAR as described in Pencina and D'Agostino (2004), Uno et al. (2011), and Gerds et al. (2013).

To account for the censoring mechanism in evaluation of predictive accuracy, we consider a working model denoted by $\mathcal{M}_2$ relating C to $\mathbf{Z}$, which is parameterized by η and represents the censoring mechanism. For example, we can use a Cox PH model for $\mathcal{M}_2$ with the conditional survival function for C as $\pi(c|\mathbf{Z}; \eta) = \exp \left\{ -\exp(\eta^T \mathbf{Z}) \int_0^c h_0(u) du \right\}$. The estimator of η , denoted by $\hat{\eta}$, can be obtained using $\mathcal{D}_{test}$ alone or using the whole data $\mathcal{D}$

if the same censoring mechanism is anticipated in both $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$. Based on $\mathcal{M}_2$, we define the condition survival function for C by $\pi(c|\mathbf{Z}; \boldsymbol{\eta}) = Pr(C > c|\mathbf{Z}; \boldsymbol{\eta})$. Under CCAR, $\pi(c|\mathbf{Z}; \boldsymbol{\eta}) = \pi(c)$. Under CAR, one estimator of truncated IPCW $\alpha(t)$ using only the pairs with both observed survival times is defined as

$$\hat{c}_g^1(\hat{\boldsymbol{\beta}}; t) = \frac{\sum_{i \neq j} \frac{\delta_i \delta_j}{\hat{\pi}_i \hat{\pi}_j} I(\tilde{S}_i < \tilde{S}_j, Y_i < Y_j, Y_i \leq t)}{\sum_{i \neq j} \frac{\delta_i \delta_j}{\hat{\pi}_i \hat{\pi}_j} I(Y_i < Y_j, Y_i \leq t)}, \quad (4)$$

where $\hat{\pi}_i = \pi(Y_i|\mathbf{Z}_i; \hat{\boldsymbol{\eta}})$ and $\hat{\pi}_j = \pi(Y_j|\mathbf{Z}_j; \hat{\boldsymbol{\eta}})$ based on the observed data from $\mathcal{D}_{test}$ or $\mathcal{D}$. Similarly, the estimator of truncated IPCW $\alpha(t)$ as the extension of $\hat{c}_g^2$ is defined as

$$\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t) = \frac{\sum_{i \neq j} \frac{\delta_i}{\hat{\pi}_i \hat{\pi}_j^*} I(\tilde{S}_i < \tilde{S}_j, Y_i < Y_j, Y_i \leq t)}{\sum_{i \neq j} \frac{\delta_i}{\hat{\pi}_i \hat{\pi}_j^*} I(Y_i < Y_j, Y_i \leq t)}, \quad (5)$$

with $\hat{\pi}_j^* = \pi(Y_j|\mathbf{Z}_j; \hat{\boldsymbol{\eta}})$. As alluded to previously, $\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t)$ has a similar format to the estimators investigated in Uno et al. (2011) and Gerds et al. (2013) but with different models for estimating censoring weights. In the following sections, we investigate several issues associated with these estimators that have not been investigated yet.

2.3 Asymptotic Properties of $\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t)$ under CAR

Theorem 2.1: Under the regularity conditions of (C1)–(C3) in Web Appendix, $\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t)$ is a consistent estimator of $\alpha(t)$, i.e., $\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t) \xrightarrow{P} c(t)$. Furthermore, $\sqrt{n} \left(\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t) - c(t) \right)$ converges weakly to a zero-mean Gaussian process.

The asymptotic properties of $\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t)$ under CAR are provided in Theorems 2.1 above. Gerds et al. (2013) reported similar results on consistency but our results on asymptotic normality are new. A sketch of the proof for Theorems 2.1 is provided in the Supplementary Materials.

The asymptotic results for $\hat{c}_g^1(\hat{\boldsymbol{\beta}}; t)$ can be derived along similar lines.

2.4 Sensitivity Analysis under NCAR

In the case of NCAR, we propose a sensitivity analysis approach, similar in spirit to the work by Siannis et al. (2005) and by Long et al. (2011). Specifically, we extend $\mathcal{M}_2$ for C to include T as a predictor, and it follows that the conditional survival function for C becomes $\pi(C|\mathbf{Z}, T; \boldsymbol{\eta}, \boldsymbol{\alpha})$ with the sensitivity parameter $\boldsymbol{\alpha}$ quantifying the association between C and T conditional on $\mathbf{Z}$. Since T is not observed for all subjects, $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ cannot be jointly estimated from the observed data. As a result, $\boldsymbol{\alpha}$ is fixed to a set of values in sensitivity analysis. Based on the extended $\mathcal{M}_2$, we obtain censoring weights as $\hat{\pi}_i = \pi(C_i|\mathbf{Z}_i, T_i; \hat{\boldsymbol{\eta}}, \boldsymbol{\alpha})$, and then compute $\hat{c}_g^1(\hat{\boldsymbol{\beta}}; t)$ and $\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t)$ for a range of pre-specified $\boldsymbol{\alpha}$ values.

To illustrate the idea more clearly, we consider the case where $\mathcal{M}_2$ is a Cox PH model with the conditional hazard function $h(c|\mathbf{Z}, T) = h_0(c) \exp(\boldsymbol{\eta}^T \mathbf{Z} + \alpha T)$. Then, $\alpha = 0$ implies CAR (i.e., $C \perp T|\mathbf{Z}$) and $\alpha \neq 0$ implies NCAR (i.e., $C \not\perp T|\mathbf{Z}$). The resulting survival function for C is $\pi(c|\mathbf{Z}, T; \boldsymbol{\eta}, \alpha) = \exp\{-\Lambda_0(c) \exp(\boldsymbol{\eta}^T \mathbf{Z} + \alpha T)\}$ with the baseline cumulative hazard $\Lambda_0(c) = \int_0^c h_0(u) du$. We propose the following estimating equation to estimate $\boldsymbol{\eta}$ for each pre-specified value α_0 :

$$U(\boldsymbol{\eta}; \alpha_0, \mathbf{Z}_i, Y_i) = \sum_{i=1}^n \left(\frac{\delta_i}{\hat{\pi}_i} - 1 \right) g(\mathbf{Z}_i) \quad (6)$$

where $\hat{\pi}_i = \hat{\pi}(Y_i|\mathbf{Z}_i, T_i; \boldsymbol{\eta}, \alpha_0) = \exp\{-\hat{\Lambda}_0(Y_i) \exp(\boldsymbol{\eta}^T \mathbf{Z}_i + \alpha_0 Y_i)\}$ and $g(\mathbf{Z}_i)$ is a known function that maps $\mathbf{Z}_i$ to $\mathcal{R}^p$ with p as the dimension of $\boldsymbol{\eta}$, e.g., $g(\mathbf{Z}_i) = \mathbf{Z}_i$, which is used in the current work. It can be shown that the set of estimation equations (6) is unbiased for any $g(\mathbf{Z}_i)$, i.e., $E\{U(\boldsymbol{\eta}, \alpha, \mathbf{Z}_i, Y_i)\} = 0$, if $\mathcal{M}_2$ is correctly specified. The estimate $\hat{\boldsymbol{\eta}}$ can be obtained using a Newton-Raphson algorithm; however, the challenge is to estimate the cumulative baseline hazard $\hat{\Lambda}_0(C_i)$. Along the lines of Scharfstein et al. (1999), we derive the modified non-parametric profile estimator adjusted to our current set-up based on the fact that $E[\delta \pi(u|\mathbf{Z}, \boldsymbol{\eta}) / \pi(Y|\mathbf{Z}, \boldsymbol{\eta}) | \mathbf{Z}, Y, C = u] = 1$. The estimation of $\boldsymbol{\eta}$ and the estimation of $\Lambda_0(C_i)$ are conducted iteratively until convergence and the algorithm is provided in Web Appendix B. We then compute $\hat{\pi}$ using $\hat{\boldsymbol{\eta}}$ obtained from the estimating equation (6).

Subsequently, the predictive accuracy estimators, $\hat{c}_g^1$ and $\hat{c}_g^2$, are obtained based on (4) and (5) respectively by plugging in $\hat{\pi}$.

2.5 High-dimensional Data

We investigate the approaches for obtaining $\hat{c}_g^1(\hat{\boldsymbol{\beta}}; t)$ and $\hat{c}_g^2(\hat{\boldsymbol{\beta}}; t)$ for $\mathcal{M}_1$ in the presence of high-dimensional data $\mathbf{Z}_i = \{Z_{i1}, Z_{i2}, \dots, Z_{ip}\}^T$ with a relatively large p or $p > n$. For the purpose of exposition, we consider the case where the censoring model $\mathcal{M}_2$ is a Cox PH model with $\pi(c|\mathbf{Z}, \boldsymbol{\eta}) = \exp\{-\Lambda_0(c) \exp(\boldsymbol{\eta}^T \mathbf{Z})\}$, and $\boldsymbol{\eta}$ can be estimated by maximizing the partial likelihood function denoted by $\mathcal{P}\mathcal{L}(\boldsymbol{\eta}, c)$ in the presence of low-dimensional data. For the high-dimensional setting, we propose to fit $\mathcal{M}_2$ using the penalization approach.

2.5.1 The Case of CAR—In the presence of high-dimensional covariates, we consider the use of the lasso and the elastic net penalties to obtain censoring weights. The penalized parameter estimates for the censoring model $\mathcal{M}_2$ are defined by

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \{-\mathcal{P}\mathcal{L}(\boldsymbol{\eta}; c) + \lambda P(\boldsymbol{\eta}; k)\} \quad (7)$$

where $P(\boldsymbol{\eta}; k) = P_{\mathcal{L}_1}(\boldsymbol{\eta}; k) + P_{\mathcal{L}_2}(\boldsymbol{\eta}; k) = k \sum_{i=1}^p |\eta_i| + (1-k) \sum_{i=1}^p \eta_i^2$, tuning parameter $\lambda > 0$ indicates the amount of penalization, and $k \in (0, 1]$ quantifies the control between sparsity and group effect. When $k = 1$, $P(\boldsymbol{\eta}, k)$ becomes the lasso penalty ($\mathcal{L}_1$). We use the R functions, `cv.glmnet` and `glmnet`, to fit the penalized Cox PH model for censoring time,

where the optimal λ is selected based on cross-validation. In particular, two approaches are considered for $\hat{\eta}$ in $\mathcal{M}_2$. In the first approach, we obtain $\hat{\eta}$ directly from the penalized Cox PH model; in the second approach, we use the penalized Cox PH model to conduct variable selection among $\mathbf{Z}$ and then fit a standard Cox PH model using the selected predictors.

Using $\hat{\eta}$, we first compute $\hat{\pi}(\mathbf{Z}, \hat{\eta})$, and subsequently $\hat{c}_g^1(\hat{\beta}; t)$ and $\hat{c}_g^2(\hat{\beta}; t)$.

2.5.2 The Case of NCAR—In the case of NCAR, it is not trivial to extend the sensitivity analysis in Section 2.4 to handle high-dimensional covariates. One intuitive two-step approach is as follows. The first step is to fit a regularized regression for the censoring model as described in Section 2.5.1 to select a subset of predictors that are potentially associated with censoring time. The second step is to conduct the sensitivity analysis proposed in Section 2.4 by utilizing only the variables selected from the first step, that is, compute $\hat{c}_g^1(\hat{\beta}; t)$ and $\hat{c}_g^2(\hat{\beta}; t)$ for a range of specified α values where the censoring weight $\hat{\pi}$ is obtained using the subset of predictors selected from the first step. This approach is straightforward and can be implemented using existing software, which is adopted for the motivating data analysis.

An alternative approach is based on penalized estimating equations (Fu, 2003; Johnson et al., 2008; Johnson, 2008), allowing for simultaneous variable selection and parameter estimation. Given the proposed estimating equation (7) with a pre-specified value of α as

α_0 , we can have $U^P(\eta; \alpha_0, \mathbf{Z}_i, Y_i) = \sum_{i=1}^n \left(\frac{\delta_i}{\pi_i} - 1 \right) g(\mathbf{Z}_i) - n\lambda \dot{P}(\eta; \alpha_0)$, where

$\dot{P}(\eta; \alpha_0) = \frac{\partial P(\eta; \alpha_0)}{\partial \eta}$ is the derivative of the penalized estimating equations with respect to η . The validity of this approach in the settings of our interest will be pursued for future research.

3. Data Analysis

The prostate cancer study includes 108 patients with a censoring rate of 61.1%. The observed failure times with the range of 2–145 months are shown in Figure 1. PSA and all gene expression values are standardized to have mean 0 and unit standard deviation, and Gleason scores are dichotomized into binary categories as less than 7 or not (33 patients with Gleason scores <7 and 75 patients with Gleason scores ≥ 7). In the exploratory analysis, the genomic biomarkers are ranked and labeled according to the score-test p -values based on univariate Cox PH models for the failure time from the lowest to the highest, which is used as a pre-screening step in our analysis (Fan and Lv, 2008; Zhao and Li, 2012). In addition, the score-test p -values for PSA and Gleason scores for the failure time are 0.13 and 0.01 respectively, and those for the censoring time are 0.87 and 0.32 respectively. Here, we consider both possible censoring mechanisms, CAR and NCAR, for assessing predictive accuracy of prediction models of interest. Of note, the Lasso Cox PH model for the censoring time select 25 genomic biomarkers with non-zero parameter estimates.

Under CAR, we evaluate five prediction models: 1) Cox PH model including PSA and Gleason (Model I); 2) Cox PH model including PSA, Gleason and biomarkers 2, 4, 8, 11, 14, 16, 22, 31, 46, 52, 63 selected in Long et al. (2011) using an earlier version of the data

(Model II); 3) Cox PH model including PSA, Gleason and biomarkers 2, 3, 5, 6, 7, 13, 15, 26, 32, 43, 56, 83 selected from all 1,536 biomarkers based on Lasso Cox PH (Model III); 4) Cox PH model including PSA, Gleason and biomarkers 2, 5, 6, 7, 10, 13, 15 selected from the top 25 biomarkers based on Lasso Cox PH (Model IV); 5) Cox PH model including PSA, Gleason and biomarkers 2, 5, 6, 7, 10 selected from the top 10 biomarkers based on Lasso Cox PH (Model V). We adopt the following procedures for assessing predictive accuracy.

1. The observed data are randomly split into two datasets, a training data $\mathcal{D}_{train}$ (50%) and a testing data $\mathcal{D}_{test}$ (50%).
2. The prediction model of interest is fitted using the $\mathcal{D}_{train}$, and then the parameter estimates are used to compute the predictive risk scores for $\mathcal{D}_{test}$.
3. The censoring model is fitted using the whole data, and the estimated parameters are used for $\mathcal{D}_{test}$ to obtain censoring weights.
4. The predictive accuracy estimated by $\hat{c}_s^2$ and $\hat{c}_g^2$ for each model are calculated using $\mathcal{D}_{test}$.

We repeat the above procedures 1,000 times, and report the average and standard deviation (SD) of the estimates for $\alpha(t)$ with t as the 25% percentile (16-month), the 75% percentile (68-month) and the maximum (164-month) of the follow-up time. The results under the assumption of CAR are summarized in Table 1. Two approaches of estimation of censoring weights are considered, Lasso-Cox1 and Lasso-Cox2, where Lasso-Cox1 estimates the censoring weights from a Cox PH model using the selected variables based on a Lasso Cox PH model (Tibshirani, 1997) for censoring time; Lasso-Cox2 utilizes censoring weights estimated directly from a Lasso Cox PH model (Zou, 2006; Yang and Zou, 2013). For all three truncated time points (16-month, 68-month and 164-month), Model III always has the highest estimates of predictive accuracy and is followed by Model IV, Model V, Model I and Model II across different methods for censoring weights. Also, the standard deviations from Model II are strictly the largest, representing the worst performance among all models. Of note, the predictive biomarkers in Model II (Long et al., 2011) were initially selected from an early version of the prostate cancer data and our estimates accounting for CAR are higher than what was reported in the previous work (Long et al., 2011). On the other hand, there still exist various degrees of differences between $\hat{c}_s^2$ and $\hat{c}_g^2$, showing the necessity to adjust for the censoring mechanism. In addition, the biomarkers are correlated with mean pairwise correlation 0.50 and median pairwise correlation 0.53; we repeat the assessment using the elastic net penalty for obtaining censoring weights, and find that Model III remains the best prediction model with the highest predicted accuracy estimates (the results are not shown).

To further assess predictive accuracy for Models I–V under NCAR, we conduct a sensitivity analysis using the first approach described in Section 2.4 to compute $\hat{c}_g^2$ with different values of the sensitivity parameter α . Based on earlier investigation, the top 100 biomarkers are included in estimation of censoring weights for sensitivity analysis, and three biomarkers are initially selected based on lasso Cox PH model. Figure 2 presents the result of sensitivity analysis with $\alpha_0 = -0.02, -0.015, -0.01, -0.005, 0, 0.005, 0.01, 0.015, 0.02$; $\hat{c}_g^2$ ranges from

0.47–0.74 for Model I, 0.41–0.58 for Model II, 0.66–0.75 for Model III, 0.47–0.72 for Model IV, and 0.41–0.65 for Model V. There is a roughly decreasing trend for all estimators as the specified sensitivity parameter α_0 increases. While the impact of α on predictive accuracy measures is not negligible, the estimated predictive accuracy for Model III is always the highest except for $\alpha_0 = 0.005$ but still close to the highest value, further lending support to the superior prediction performance of Model III over the other alternative models.

4. Simulation Studies

We conduct extensive simulations to evaluate the four estimators of the predictive metric $\alpha(t)$, namely, $\hat{c}_s^1$, $\hat{c}_s^2$, $\hat{c}_g^1$ and $\hat{c}_g^2$ in settings with low-dimensional and high-dimensional covariates under CAR and NCAR. To benchmark the bias of the four estimators, we compute the “true” predictive accuracy metric with a large sample size of 1, 000, 000 given by

$$c_{true}(\beta_0; t) = \frac{\sum_{i \neq j} I(S_i < S_j, T_i < T_j, T_i \leq t)}{\sum_{i \neq j} I(T_i < T_j, T_i \leq t)} \quad (8)$$

where the underlying true failure times and the calculated predicted scores S using the true parameter β_0 are used. Note that $c_{true}(\beta_0; t)$ cannot be calculated in real data. The failure time T and the censoring time C are generated from Cox-exponential models (Bender et al., 2005). Next, we will focus on the cases of CAR, and the simulation settings and corresponding results under NCAR are provided in the Supplementary Materials.

First, in the settings with low-dimensional covariates, $\mathbf{Z} = (Z_1, Z_2, Z_3)^T$, T and C are assumed to follow exponential distributions with hazard functions given by

$$\begin{aligned} h(T|\mathbf{Z}) &= \lambda_{0T} \exp(\boldsymbol{\beta}^T \mathbf{Z}) = \lambda_{0T} \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3) \\ h(C|\mathbf{Z}) &= \lambda_{0C} \exp(\boldsymbol{\eta}^T \mathbf{Z}) = \lambda_{0C} \exp(\eta_1 Z_1 + \eta_2 Z_2 + \eta_3 Z_3) \end{aligned} \quad (9)$$

where the constant baseline hazards λ_{0T} set as 0.1 or 0.6 and $\lambda_{0C} = 0.1$, the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T = (-1, 0.5, 2)^T$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)^T = (1, 1, 1)^T$, and the covariates $Z_1 \sim \text{Unif}[0, 1]$, $Z_2 \sim \text{Unif}[-1, 1]$ and $Z_3 \sim N(0, 1)$. The event rates denoted by ζ are around 0.3 ($\lambda_{0T} = 0.1$) and 0.6 ($\lambda_{0T} = 0.6$). We evaluate two prediction models, the correctly specified Cox PH model including all three covariates of Z_1 , Z_2 and Z_3 and an incorrectly specified AFT model with the log-logistic distribution including all three covariates of Z_1 , Z_2 and Z_3 . Hence, there are four combinations of the prediction model for survival time ($\mathcal{M}_1$) and the working model for censoring time ($\mathcal{M}_2$), namely, Cox($\mathcal{M}_1$)-Cox($\mathcal{M}_2$), Cox($\mathcal{M}_1$)-AFT($\mathcal{M}_2$), AFT($\mathcal{M}_1$)-Cox($\mathcal{M}_2$), and AFT($\mathcal{M}_1$)-AFT($\mathcal{M}_2$). The impact of mis-specification of the censoring models on predictive accuracy estimation will be investigated.

Second, in the settings with $p = 10$ or 100 covariates, $\mathbf{Z}$, T and C are also generated from exponential distributions with the following hazard functions

$$\begin{aligned} h(T|\mathbf{Z}) &= \lambda_{0T} \exp(\boldsymbol{\beta}^T \mathbf{Z}) = \lambda_{0T} \exp(\beta_1 Z_1 + 0 \times Z_2 + \beta_3 Z_3 + 0 \times Z_4 + \beta_5 Z_5 + \dots) \\ h(C|\mathbf{Z}) &= \lambda_{0C} \exp(\boldsymbol{\eta}^T \mathbf{Z}) = \lambda_{0C} \exp(\eta_1 Z_1 + 0 \times Z_2 + \eta_3 Z_3 + 0 \times Z_4 + \eta_5 Z_5 + \dots) \end{aligned} \quad (10)$$

where the constant baseline hazards λ_{0T} as 0.05 or 2 and $\lambda_{0C} = 0.2$, and the parameters $\boldsymbol{\beta} = (-1, 0, 0.5, 0, 2, 0, \dots, 0)^T$ and $\boldsymbol{\eta} = (1, 0, 1, 0, 1, 0, \dots, 0)^T$. The covariates $\mathbf{Z}$ are generated from a multivariate standard normal distribution $MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an identity matrix (independence among predictors) or an exchangeable matrix with a correlation coefficient $\rho = 0.8$ (high pairwise-correlation among predictors) only in the setting with $p = 100$. The event rates ζ are about 0.3 ($\lambda_{0T} = 0.05$) and 0.8 ($\lambda_{0T} = 2$). In these settings, we evaluate the prediction performance of the correctly specified Cox PH model including Z_1 , Z_3 and Z_5 .

Each simulated data set is equally divided into two subsets $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ with sample size $n = 50, 100, 500$. The observed survival time is defined as $Y_i = \min(T_i, C_i)$ and the event indicator is defined as $\delta_i = I(T_i < C_i)$ ($i = 1, 2, \dots, n$). The estimated prediction models are obtained from $\mathcal{D}_{train}$, and the predictive scores are calculated from $\mathcal{D}_{test}$. The censoring weights are estimated using the whole data $\mathcal{D}$ because the censoring mechanism is the same in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ which are obtained by random splitting. In particular, in the first scenario, all three covariates $\mathbf{Z} = (Z_1, Z_2, Z_3)^T$ are included in the Cox PH model or AFT model for the censoring times; while for the second scenario, three approaches are used to obtain the censoring weights, Cox representing that the censoring weights are obtained from a Cox PH model, and Lasso-Cox1 and Lasso-Cox2 are the same as describe in Section 3. Thereafter, we calculate the four predictive accuracy estimators defined in Section 2 and their differences compared with $c_{true}(\boldsymbol{\beta}_0; t)$. Note that the truncated time t is chosen as 20, which is above 75% percentile of the failure times. To summarize simulation results, 1,000 Monte Carlo data sets are generated for each setting. For each estimator, we report the absolute value of Monte Carlo mean bias denoted by $\Delta \hat{c}_s^1$, $\Delta \hat{c}_s^2$, $\Delta \hat{c}_g^1$ and $\Delta \hat{c}_g^2$ with Monte Carlo SD in parentheses.

Table 2 presents the simulation results for the setting with $p = 3$. In all cases, $\hat{c}_g^2$ performs the best due to the smallest bias among all the estimators. In particular, when the censoring model is correctly specified, $\Delta \hat{c}_g^2$ is negligible. $\hat{c}_g^2$ is fairly robust against the misspecification of the censoring model as an AFT model when the censoring rate is not high. $\Delta \hat{c}_g^2$ becomes slightly larger as the censoring rate and sample size increase. In addition, $\hat{c}_s^1$ and $\hat{c}_g^1$ are less efficient than $\hat{c}_s^2$ and $\hat{c}_g^2$, likely as a result of $\hat{c}_s^1$ and $\hat{c}_g^1$ using less information compared to $\hat{c}_s^2$ and $\hat{c}_g^2$; $\Delta \hat{c}_g^1$ and $\Delta \hat{c}_g^2$ are smaller than $\Delta \hat{c}_s^1$ and $\Delta \hat{c}_s^2$ respectively, likely as a result of $\hat{c}_g^1$ and $\hat{c}_g^2$ using information from the censoring model in estimation.

Table 3 shows the simulation results for the settings with $p = 10$. When the approaches of Cox and Lasso-Cox1 are used to obtain the censoring weights, $\hat{c}_g^2$ performs better than the other estimators, followed by $\hat{c}_s^2$, $\hat{c}_g^1$ and $\hat{c}_s^1$, and its bias is negligible. The standard deviation of $\Delta\hat{c}_g^2$ is slightly larger than $\Delta\hat{c}_s^2$ when the event rate is 0.3, but the difference attenuates towards 0 as the event rate increases to 0.8; unsurprisingly, the standard deviation of $\Delta\hat{c}_s^2$ and $\Delta\hat{c}_g^2$ are lower than $\Delta\hat{c}_s^1$ and $\Delta\hat{c}_g^1$ respectively. In addition, $\hat{c}_g^2$ achieves similar performance when the approaches of Cox and Lasso-Cox1 are used to obtain the censoring weights but its performance deteriorates when the approach of Lasso-Cox2 is used, suggesting that the approaches of Cox and Lasso-Cox1 are preferred in these settings.

Table 4 presents the simulation results with focus on IPCW c -statistic for the moderately high-dimensional setting with $p = 100$ covariates and $n = 50$ or 100 . Two correlation structures are considered for the covariates, independence with $\rho = 0$ and pairwise correlation with $\rho = 0.8$. In these settings, the approach of Cox for estimating censoring weights is no longer applicable, thus the approaches of Lasso-Cox1 and Lasso-Cox2 are used. For the cases with $\rho = 0.8$, the elastic net is also considered for obtaining censoring weights to deal with high correlation among high-dimensional predictors, where Elastic-Net-Cox1 uses a Cox PH model including only covariates selected by an Elastic-net Cox PH model for censoring time with the weight of lasso versus ridge as $k = 0.2$ and Elastic-Net-Cox2 estimates censoring weights directly using an Elastic-net Cox PH model for censoring time. Similar to the findings in Table 3, for the scenarios with $\rho = 0$, $\Delta\hat{c}_g^2$ using the approach of Lasso-Cox2 is larger compared to the approach of Lasso-Cox1, and the difference becomes less pronounced as the event rate increases. Still, $\hat{c}_g^2$ based on Lasso-Cox1 shows negligible bias and achieves the best performance. In the presence of high correlation (i.e., $\rho = 0.8$) among predictors, Elastic-Net-Cox1 performs the best evidenced by the smallest values of $\Delta\hat{c}_g^1$ and $\Delta\hat{c}_g^2$ compared with the other approaches for censoring weights, but Lasso-Cox1 also performs satisfactory because of its small or negligible bias when the sample size is as large as 100. Interestingly, when the event rate is high (i.e., $\zeta = 0.8$) and the sample size is 100, all estimates, based on lasso or elastic net, are comparable in terms of bias. The results in Table 4 show that Lasso-Cox1 achieves the best or close to the best performance for $n = 100$ even when the correlation among variables is high. The simulation results for the cases with $p = 1,000$ (not shown here) are similar to Table 4.

Additional simulations are conducted to investigate the impacts of NCAR on predictive accuracy metric estimation for the settings with low- and high-dimensional covariates, focusing on $\hat{c}_g^2$; the detailed information is provided in Appendix C of the Supplementary Materials. Our simulation results in these settings show that the bias $\Delta\hat{c}_g^2$ increases as the sensitivity parameter α moves away from its true value, particularly, when $\alpha = 0$ (i.e. CAR). For the same α value, the magnitude of biases seems to be comparable in low-dimensional and high-dimensional scenarios and higher event rates (i.e. $\zeta = 0.6$) lead to decreased biases.

5. Discussion

Based on our results, we recommend using $\hat{c}_g^2$ and conducting sensitivity analyses to assess the potential impacts of the NCAR assumption for both low-dimensional and high-dimensional data. In the data example, the sensitivity analysis shows that the impact of the sensitivity parameter α is not negligible. In addition, our simulations show that the assumption of CAR can lead to biased estimation of predictive accuracy if CAR does not hold; the magnitude of biases, though, seems to be comparable under the settings of low-and high-dimensionality. Under CAR, $\hat{c}_g^2$ achieves the best performance among the estimators considered in settings with low-dimensional or high-dimensional data. Also, $\hat{c}_g^2$ is fairly robust to the mis-specification of the censoring model when the censoring rate is not high. In the presence of high-dimensional data, the approach of Lasso-Cox1, which obtains censoring weights from a Cox PH model including covariates selected by a Lasso Cox PH model for the censoring time, achieves satisfactory performance with small to negligible bias even in the presence of high correlation.

There are several topics for future research. First, in the presence of high-dimensional data under the NCAR assumption, the proposed sensitivity analysis based on regularized estimating equations for survival endpoints needs further investigation. Second, it is of potential interest to investigate time-dependent IPCW c -statistic under the NCAR assumption. Third, the proposed approaches for assessing prediction models with survival outcomes can be extended to other predictive accuracy metrics such as net reclassification index and integrated discrimination index.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Wang's research was partially supported by the start-up funding from the Department of Public Health Sciences at Pennsylvania State Hershey Medical Center and a KL2 career grant supported by the National Center for Advancing Translational Sciences, Grant KL2 TR000126. Long's research was partially supported by NIH/NCI grants R03CA173770 and R03CA183006.

References

- Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazard models. *Statistics in Medicine*. 2005; 24:1713–1723. [PubMed: 15724232]
- Cox D. Regression models and life-tables. *Journal of the Royal Statistical Society Series B*. 1972; 34:187–202.
- D'Agostino R, Griffith J, Schmidt C, Terrin N. Measures of evaluating model performance. *Proceedings of the Biometrics Section, American Statistical Association*. 1997:253–258.
- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*. 2008:849–911.
- Freedman A, Seminara D, Gail M, Hartge P, Colditz G, Ballard-Barbash R, Pfeiffer R. Cancer risk prediction models: a workshop on development, evaluation, and application. *Journal of the National Cancer Institute*. 2005; 97:715–723. [PubMed: 15900041]

- Freedman A, Slattery M, Ballard-Barbash R, Willis G, Cann B, Pee D, Gail M, Pfeiffer R. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *Journal of Clinical Oncology*. 2009; 27:686–693. [PubMed: 19114701]
- Fu W. Penalized estimating equations. *Biometrics*. 2003; 59:126–132. [PubMed: 12762449]
- Gerds T, Kattan M, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*. 2013; 32:2173–2184. [PubMed: 23172755]
- Gonen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005; 92:965–970.
- Harrell F, Lee K, Mark D. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996; 15:361–387. [PubMed: 8668867]
- Heagerty P, Zheng Y. Survival model predictive accuracy and roc curves. *Biometrics*. 2005:92–105. [PubMed: 15737082]
- Heitjan D, Rubin D. Ignorability and coarse data. *The Annals of Statistics*. 1991; 19:1681–2284.
- Hoerl A, Kennard R. Ridge regression: Applications to nonorthogonal problems. *Technometrics*. 1970; 12:69–82.
- Johnson BA. Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70:351–370.
- Johnson BA, Lin D, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*. 2008; 103:672–680. [PubMed: 20376193]
- Long Q, Chung M, Moreno C, Johnson B. Risk prediction for prostate cancer recurrence through regularized estimation with simultaneous adjustment for nonlinear clinical effects. *Annals of Applied Statistics*. 2011; 5:2003–2023. [PubMed: 22081781]
- Long Q, Johnson B, Osunkoya A, Lai Y, Zhou W, Abramovitz M, Xia M, Bouzyk M, Nam R, Sugar L, Stanimirovi A, Leyland-Jones B, Seth A, Petros J, Moreno C. Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy. *American Journal of Pathology*. 2011; 179:46–54. [PubMed: 21703393]
- Long Q, Zhang X, Johnson BA. Robust estimation of area under roc curve using auxiliary variables in the presence of missing biomarker values. *Biometrics*. 2011; 67:559–567. [PubMed: 20825391]
- Optenberg S, Clark J, Brawer M, Thompson I, Stein C, Friedrichs P. Development of a decision-making tool to predict risk of prostate cancer: the cancer of the prostate risk index (capri) test. *Urology*. 1997; 50:665–672. [PubMed: 9372872]
- Pencina MJ, D'Agostino RB. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*. 2004; 23:2109–2123. [PubMed: 15211606]
- Scharfstein D, Rotnitzky A, Robins J. Adjusting for nonignorable dropout using semiparametric nonresponse models. *Journal of the American Statistical Association*. 1999; 94:1135–1146.
- Siannis F, Copas J, Lu G. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*. 2005; 6:77–91. [PubMed: 15618529]
- Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, Pencina M, Kattan M. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010; 21:128–138. [PubMed: 20010215]
- Tibshirani R. The lasso method for variable selection in the cox model. *Statistics in Medicine*. 1997; 16:385–395. [PubMed: 9044528]
- Uno H, Cai T, Pencina M, D'Agostino R, Wei J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*. 2011; 30:1105–1117. [PubMed: 21484848]
- Yang Y, Zou H. A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions statistics and its interface. *Statistics and Its Interface*. 2013:167–173.
- Zeng D, et al. Estimating marginal survival function by adjusting for dependent censoring using many covariates. *The Annals of Statistics*. 2004; 32:1533–1555.

- Zhao SD, Li Y. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*. 2012;397–411. [PubMed: 22408278]
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*. 2005; 67:301–320.

Figure 1.
The KM curves for the failure time and censoring time in the prostate cancer study

Figure 2.
Sensitivity analysis for the Prostate Cancer study.

Table 1

Summary of predictive accuracy estimates for five models of interest. Lasso-Cox1: censoring weights are estimated from a Cox PH model using the selected variables based on a Lasso-Coxph model. Lasso-Cox2: censoring weights are estimated directly from a Lasso-Cox model.

Model	$\hat{c}_s^2(SD)$	$\hat{c}_g^2(SD)$	$\hat{c}_s^2(SD)$	$\hat{c}_g^2(SD)$	$\hat{c}_s^2(SD)$	$\hat{c}_g^2(SD)$	$t_{MAX} \leftarrow 164(\text{month})$
	$t_{Q25} = 16(\text{month})$		$t_{Q75} = 68(\text{month})$		Lasso-Cox1		
Model I	0.599(0.053)	0.599(0.054)	0.587(0.049)	0.590(0.052)	0.588(0.050)	0.589(0.056)	
Model II	0.561(0.097)	0.561(0.097)	0.540(0.068)	0.541(0.068)	0.539(0.065)	0.543(0.068)	
Model III	0.723(0.072)	0.723(0.071)	0.705(0.056)	0.707(0.057)	0.704(0.055)	0.710(0.058)	
Model IV	0.683(0.070)	0.680(0.071)	0.666(0.055)	0.672(0.057)	0.666(0.054)	0.668(0.059)	
Model V	0.670(0.075)	0.665(0.076)	0.638(0.058)	0.638(0.060)	0.641(0.057)	0.638(0.061)	
			Lasso-Cox2				
Model I	0.599(0.053)	0.577(0.103)	0.587(0.049)	0.545(0.084)	0.588(0.050)	0.546(0.084)	
Model II	0.561(0.097)	0.550(0.124)	0.540(0.068)	0.528(0.098)	0.539(0.065)	0.528(0.098)	
Model III	0.723(0.072)	0.726(0.095)	0.705(0.056)	0.715(0.081)	0.704(0.055)	0.715(0.081)	
Model IV	0.683(0.070)	0.669(0.101)	0.666(0.055)	0.663(0.080)	0.666(0.054)	0.663(0.080)	
Model V	0.670(0.075)	0.600(0.116)	0.638(0.058)	0.579(0.089)	0.641(0.057)	0.580(0.088)	

Model I: only PSA and Gleason;

Model II: PSA, Gleason and biomarkers 2,4,8,11,14,16,22,31,46,52,63 based on Long et al. (2011);

Model III: PSA, Gleason and biomarkers 2,3,5,6,7,13,15,26,32,43,56,83 from all 1536 biomarkers;

Model IV: PSA, Gleason and biomarkers 2,5,6,7,10,13,15 from the top 25 biomarkers;

Model V: PSA, Gleason and biomarkers 2,5,6,7,10 from the top 10 biomarkers;

Table 2

Simulation results for the settings with $p = 3$ and $t = 50$.

$\mathcal{M}$	ζ	n	$\mathcal{M}_2$	$\Delta \hat{c}_s^1(SD)$	$\Delta \hat{c}_s^2(SD)$	$\Delta \hat{c}_g^1(SD)$	$\Delta \hat{c}_g^2(SD)$	c_{true}
Cox	0.3	50	Cox	0.076(0.072)	0.020(0.049)	0.062(0.092)	0.000(0.062)	0.834
			AFT [‡]			0.069(0.075)	0.003(0.050)	
		500	Cox	0.068(0.020)	0.027(0.014)	0.040(0.040)	0.003(0.015)	
			AFT [‡]			0.054(0.019)	0.011(0.013)	
	0.6	50	Cox	0.035(0.042)	0.010(0.035)	0.023(0.045)	0.001(0.034)	0.830
			AFT [‡]			0.029(0.042)	0.002(0.034)	
		500	Cox	0.031(0.012)	0.013(0.010)	0.010(0.016)	0.001(0.010)	
			AFT [‡]			0.020(0.012)	0.000(0.010)	
AFT [‡]	0.3	50	Cox	0.076(0.072)	0.020(0.049)	0.063(0.092)	0.000(0.061)	0.834
			AFT [‡]			0.069(0.076)	0.003(0.050)	
		500	Cox	0.068(0.020)	0.027(0.014)	0.041(0.040)	0.003(0.027)	
			AFT [‡]			0.055(0.020)	0.010(0.009)	
	0.6	50	Cox	0.035(0.042)	0.010(0.034)	0.024(0.046)	0.001(0.034)	0.830
			AFT [‡]			0.029(0.042)	0.002(0.033)	
		500	Cox	0.031(0.012)	0.013(0.011)	0.010(0.016)	0.001(0.010)	
			AFT [‡]			0.020(0.012)	0.000(0.010)	

Note that both true underlying models for survival time and censoring time are Cox PH models; $\Delta \hat{c}_s^1(SD)$ and $\Delta \hat{c}_g^2(SD)$ are the same when $\mathcal{M}_2$ is Cox or AFT because they do not require the estimation of censoring weights.

[‡]Mis-specifying the model for survival time as an AFT model with the log-logistic distribution.

[‡]Mis-specifying the model for censoring time as an AFT model with the log-logistic distribution.

Table 3

Simulation results for the settings with $p = 10$ and $t = 20$. Cox: censoring weights are estimated from a Cox PH model. Lasso-Cox1: censoring weights are estimated from a Cox PH model using the selected variables based on a Lasso Cox PH model. Lasso-Cox2: censoring weights are estimated directly from a Lasso Cox PH model.

ζ	n	m_2	$\Delta \hat{c}_g^1(SD)$	$\Delta \hat{c}_g^2(SD)$	$\Delta \hat{c}_g^1(SD)$	$\Delta \hat{c}_g^2(SD)$	c_{true}
0.3	50	Cox	0.088(0.074)	0.021(0.046)	0.055(0.095)	0.002(0.056)	0.858
		Lasso-Cox1			0.051(0.093)	0.002(0.054)	
	100	Cox	0.082(0.046)	0.024(0.029)	0.102(0.082)	0.029(0.050)	
		Lasso-Cox1			0.034(0.064)	0.005(0.035)	
0.8	50	Cox	0.077(0.020)	0.027(0.013)	0.029(0.062)	0.006(0.032)	
		Lasso-Cox1			0.097(0.052)	0.032(0.033)	
	100	Cox	0.026(0.034)	0.005(0.029)	0.016(0.031)	0.006(0.015)	0.846
		Lasso-Cox1			0.013(0.031)	0.008(0.014)	
	50	Cox	0.024(0.023)	0.007(0.021)	0.092(0.022)	0.036(0.014)	
		Lasso-Cox1			0.009(0.033)	0.002(0.028)	
	100	Cox	0.022(0.010)	0.008(0.009)	0.040(0.048)	0.010(0.049)	
		Lasso-Cox1			0.004(0.022)	0.000(0.019)	
	500	Cox	0.022(0.010)	0.008(0.009)	0.004(0.023)	0.001(0.020)	
		Lasso-Cox1			0.036(0.028)	0.011(0.036)	
	500	Cox	0.022(0.010)	0.008(0.009)	0.002(0.010)	0.000(0.008)	
		Lasso-Cox1			0.003(0.010)	0.001(0.008)	
	500	Cox	0.022(0.010)	0.008(0.009)	0.033(0.011)	0.013(0.009)	
		Lasso-Cox1					

Note that $\Delta \hat{c}_g^1(SD)$ and $\Delta \hat{c}_g^2(SD)$ are the same for Cox, Lasso-Cox1 and Lasso-Cox2 because they do not require the estimation of censoring weights.

Table 4

Simulation results for settings with $p = 100$ and $t = 20$. $\mathbf{Z}$ has an exchangeable correlation structure with ρ . Lasso-Cox1: censoring weights are estimated from a Cox PH model using the selected variables based on a Lasso Cox PH model; Lasso-Cox2: censoring weights are estimated directly from a Lasso Cox PH model; Elastic-Net-Cox1: censoring weights are estimated from a Cox PH model using the selected variables based on an Elastic-net Cox PH model; Elastic-Net-Cox2: censoring weights are estimated directly from an Elastic-net Cox PH model.

ρ	ζ	n	$\mathcal{M}_2$	$\Delta \hat{c}_g^1(SD)$	$\Delta \hat{c}_g^2(SD)$	c_{me}
0	0.3	50	Lasso-Cox1	0.059(0.092)	0.003(0.062)	0.860
			Lasso-Cox2	0.107(0.085)	0.024(0.053)	
	100	100	Lasso-Cox1	0.037(0.064)	0.003(0.041)	
			Lasso-Cox2	0.099(0.054)	0.031(0.034)	
0.8	50	50	Lasso-Cox1	0.014(0.035)	0.002(0.040)	0.847
			Lasso-Cox2	0.046(0.090)	0.017(0.086)	
	100	100	Lasso-Cox1	0.007(0.023)	0.001(0.020)	
			Lasso-Cox2	0.042(0.063)	0.013(0.057)	
0.8	0.3	50	Lasso-Cox1	0.009(0.049)	0.008(0.052)	0.915
			Lasso-Cox2	0.018(0.054)	0.006(0.050)	
		100	Elastic-Net-Cox1	0.006(0.050)	0.003(0.053)	
			Elastic-Net-Cox2	0.018(0.054)	0.007(0.047)	
		100	Lasso-Cox1	0.012(0.033)	0.006(0.035)	
			Lasso-Cox2	0.016(0.034)	0.012(0.028)	
	0.8	50	Elastic-Net-Cox1	0.009(0.035)	0.002(0.037)	
			Elastic-Net-Cox2	0.016(0.034)	0.011(0.028)	
		100	Lasso-Cox1	0.004(0.027)	0.012(0.027)	0.903
			Lasso-Cox2	0.010(0.033)	0.002(0.059)	
		100	Elastic-Net-Cox1	0.004(0.026)	0.002(0.030)	
			Elastic-Net-Cox2	0.010(0.028)	0.002(0.050)	
	100	100	Lasso-Cox1	0.003(0.018)	0.009(0.019)	
			Lasso-Cox2	0.009(0.019)	0.004(0.020)	
	100	100	Elastic-Net-Cox1	0.003(0.018)	0.003(0.019)	
			Elastic-Net-Cox2	0.009(0.019)	0.004(0.021)	