Biostatistics (2016), **17**, 4, *pp.* 692–707
doi:10.1093/biostatistics/kxw016
Advance Access publication on April 22, 2016

Inference for survival prediction under the regularized Cox model

JENNIFER A. SINNOTT*
Department of Statistics, The Ohio State University, Columbus, OH 43210, USA
jsinnott@stat.osu.edu
TIANXI CAI
Department of Biostatistics, Harvard University, Boston, MA 02115, USA

SUMMARY

When a moderate number of potential predictors are available and a survival model is fit with regularization to achieve variable selection, providing accurate inference on the predicted survival can be challenging. We investigate inference on the predicted survival estimated after fitting a Cox model under regularization guaranteeing the oracle property. We demonstrate that existing asymptotic formulas for the standard errors of the coefficients tend to underestimate the variability for some coefficients, while typical resampling such as the bootstrap tends to overestimate it; these approaches can both lead to inaccurate variance estimation for predicted survival functions. We propose a two-stage adaptation of a resampling approach that brings the estimated error in line with the truth. In stage 1, we estimate the coefficients in the observed data set and in B resampled data sets, and allow the resampled coefficient estimates to vote on whether each coefficient should be 0. For those coefficients voted as zero, we set both the point and interval estimates to $\{0\}$. In stage 2, to make inference about coefficients not voted as zero in stage 1, we refit the penalized model in the observed data and in the B resampled data sets with only variables corresponding to those coefficients. We demonstrate that ensemble voting-based point and interval estimators of the coefficients perform well in finite samples, and prove that the point estimator maintains the oracle property. We extend this approach to derive inference procedures for survival functions and demonstrate that our proposed interval estimation procedures substantially outperform estimators based on asymptotic inference or standard bootstrap. We further illustrate our proposed procedures to predict breast cancer survival in a gene expression study.

Keywords: Bootstrap; Ensemble methods; Oracle property; Proportional hazards model; Regularized estimation; Resampling; Risk prediction; Simultaneous confidence intervals; Survival functions.

1. INTRODUCTION

Many modern medical studies seek to use genomic measurements to predict survival. With a small number of predictors, the standard Cox proportional hazards model (Cox, 1972) can be used to effectively make inference about survival functions. When many potential predictors are available, it is often desirable to

*To whom correspondence should be addressed.

build accurate yet parsimonious models that only use a small number of biomarkers. When the number of predictors is moderate, an approach using shrinkage to perform simultaneous variable selection and estimation, such as the lasso, can be effective (Tibshirani, 1996, 1997). Asymptotically, the lasso-penalized estimator is not consistent in variable selection and has non-regular asymptotic distributions, which results in difficulty constructing valid confidence intervals (CIs) (Knight and Fu, 2000). Several alternative penalty functions have been proposed that do possess the so-called oracle properties, in that they are consistent for variable selection and yield estimators with asymptotic normality. These penalty functions, including the adaptive lasso, the smoothly clipped absolute deviation (SCAD), and the adaptive elastic net (aENET), have been adapted to the Cox model (Fan and Li, 2002; Zou, 2006; Zhang and Lu, 2007; Zou and Zhang, 2009; Wu, 2012). There are benefits to each of these approaches; for example, the aENET has good estimation and variable selection performance in situations with correlated predictors whose effects are sparse.

For the regression parameters, denoted by $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0p})^{\mathrm{T}}$, standard error (SE) formulas for the estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^{\mathrm{T}}$ based on asymptotic results have been proposed for the adaptive lasso and SCAD, and could be analogously derived for the aENET (Fan and Li, 2002; Zhang and Lu, 2007). The basis for the derivation of these formulas is the oracle property, which tells us that the penalized estimator is asymptotically equivalent to the oracle estimator, the unpenalized estimator fit with only the "true" signals. The formulas thus rely in part on the accuracy of the variable selection achieved by the penalization: they provide non-trivial SE estimates for $\hat{\theta}_j$ when $\hat{\theta}_j \neq 0$, but set the SE to zero when $\hat{\theta}_j = 0$. This tends to yield accurate SE estimates for non-zero θ_{0j}'s, but underestimates the SE of $\hat{\boldsymbol{\theta}}_j$ when $\theta_{0j} = 0$. An alternative approach would be to obtain variance estimates with commonly used resampling methods such as the bootstrap. Unfortunately, standard resampling methods tend to overestimate the variance when the true coefficient is 0, even when the sample size is relatively large. In this paper, we first propose an ensemble voting-based procedure, an adaptation of resampling leveraging the oracle property, to provide accurate point and interval estimates for both zero and non-zero coefficients. Building on top of the ensemble procedure for coefficient estimation, we then propose resampling procedures for making precise inference about predicted survival functions at any given predictor level.

Specifically, our proposed method proceeds in two stages. In stage 1, we fit the penalized model in the observed data set and across resampled data sets, and use this collection of estimated coefficients to vote on which variables belong in the model. For each coefficient θ_j, we determine whether the proportion of $\{\hat{\theta}_j^{*(1)}, \ldots, \hat{\theta}_j^{*(B)}\}$ which are 0 is higher than a specified fraction $\mathfrak{p}_j$ and if so, we set both the point and interval estimate for θ_j to be $\{0\}$. To make inference about coefficients voted as non-zero in stage 1, in stage 2, we refit the model in both the original and resampled data sets with only those surviving variables. The refit estimates are then used to construct point and interval estimates for these coefficients. This ensemble voting-based method can be viewed as a compromise between making inferences based on the oracle property and resampling. Those voted as zero in stage 1 are deemed as "confidently zero" and hence the oracle property is applied to make inference for these coefficients. Resampling is then used to make inference about the remaining coefficients. Note that our proposed point estimator resembles the relaxed lasso estimator (Meinshausen, 2007), with one main difference being that our method determines the active set based on voting. As shown in numerical studies, the new point estimate does not differ dramatically from the initial aENET estimate, yet it allows the resampling to more accurately capture its variability and hence leads to more precise inference.

Our interest lies in inference not only on the coefficients, but also on functions of the coefficients—specifically, the predicted survival function for new patients. With regularized estimation of the regression coefficients, proper inference procedures for the survival function are not currently available. Naively making inference based on asymptotics or the bootstrap can lead to imprecise interval estimation for the survival functions. We propose to construct point and interval estimates for the survival functions building on top of the two-stage procedure for coefficient estimation and resampling. Our procedure, benefiting

from more accurate inference for the zero coefficients, can produce pointwise and simultaneous CIs with better finite sample performance than those obtained from naive methods.

Our proposed approach shares a number of features with other recent ensemble-based approaches developed for variable selection. For example, for linear models, the randomized lasso with stability selection (Meinshausen and Bühlmann, 2010) and the bootstrap lasso (Bach, 2008) both fit lasso-type procedures in observed and resampled data, and look across the resampled estimators to identify which variables should and should not be included in the model. They establish results about the consistency of variable selection guaranteed by these approaches, even when the number of potential predictors is quite large. The idea behind our proposed point estimator is similar to what is proposed in these papers, and it does inherit oracle properties from the penalized estimators it uses. However, our goal is to use an ensemble-type approach to produce both a good point estimate *and* a good collection of resampled estimators that accurately capture the variability of the point estimate in finite samples. This joint goal distinguishes our method from previous work. Furthermore, no existing methods consider downstream inference for survival functions in the presence of regularization for coefficient estimation.

The rest of the paper is organized as follows. In Section 2, we introduce our ensemble voting-based procedure, with the main methodological details provided in Section 2.2 and notes on implementation in Section 2.4. In Section 3.1, we evaluate our method using simulation studies and in Section 3.2, we demonstrate its usage for predicting the probability of breast cancer progression using a set of genes in a candidate pathway. In Section 4, we make some final comments and further situate our method in the context of other existing ensemble approaches.

2. METHODS

We consider the setting in which we have a collection of p_Z novel genomic or biological predictors $\mathbf{Z}$, and wish to use them along with p_D clinical covariates $\mathbf{D}$ to predict patient survival time T. Due to censoring, we only observe $X = \min(T, C)$ and $\Delta = I(T \leqslant C)$, where C is the censoring time assumed independent of T given $\mathbf{W} = (\mathbf{D}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$. The observed data consist of n independent and identically distributed (iid) random vectors, $\mathcal{O} = \{(X_i, \Delta_i, \mathbf{W}_i^{\mathrm{T}})^{\mathrm{T}}\}_{i=1,\ldots,n}$. Without loss of generality, we assume that Z_j's are standardized to have mean 0 and variance 1. We further assume that p_D is small and all clinical variables are included in the model; however, p_Z may be of moderate size relative to n. We assume a Cox proportional hazards model for $T \mid \mathbf{W}$, $S(t; \mathbf{W}) \equiv P(T \geqslant t \mid \mathbf{W}) = g\{\log \Lambda_0(t) + \boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{W}\}$, where $g(x) = \exp\{-\exp(x)\}$, $\Lambda_0(\cdot)$ is the unknown baseline cumulative hazard function, and $\boldsymbol{\theta}_0$ are the unknown log hazard ratio parameters. We let $\mathcal{A}^c = \{j : \theta_{0j} = 0, p_D < j \leqslant p\}$ denote the non-active set of the coefficients for $\mathbf{Z}$ and let $\mathcal{A} = \{1, \ldots, p\} \backslash \mathcal{A}^c$, where $p = p_D + p_Z$.

2.1 *Regularized estimation and initial perturbation*

Since p_Z is not small and the coefficient vector may be sparse, we may estimate $\boldsymbol{\theta}_0$ by maximizing a penalized log partial likelihood with a penalty providing simultaneous variable selection and estimation. For clarity, we will use the aENET penalty throughout but identical methods could be pursued using any penalization with oracle properties. Specifically, we focus on the estimator

$$\hat{\boldsymbol{\theta}} = (1 + \lambda_2) \operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ -\hat{\ell}_0(\boldsymbol{\theta}) + \lambda_2 \sum_{j=p_D+1}^{p} \theta_j^2 + \lambda_1 \sum_{j=p_D+1}^{p} \hat{w}_j |\theta_j| \right\},$$

where $\hat{\ell}_0(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \Delta_i [\boldsymbol{\theta}^{\mathrm{T}}\mathbf{W}_i - \log\{n \hat{\Pi}^{(0)}(\boldsymbol{\theta}, X_i)\}]$ is the log partial likelihood, $\hat{\Pi}^{(0)}(\boldsymbol{\theta}, s) = n^{-1} \sum_{i=1}^n \mathbf{I}(X_i \geqslant s) \exp(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{W}_i)$, $\hat{w}_j = |\tilde{\theta}_{Rj}|^{-1}$, and $\tilde{\boldsymbol{\theta}}_R = (1 + \lambda_2) \operatorname{argmin}_{\boldsymbol{\theta}} \{-\hat{\ell}_0(\boldsymbol{\theta}) + \lambda_2 \sum_{j=p_D+1}^{p} \theta_j^2\}$.

Here λ_1 and λ_2 are non-negative tuning parameters controlling the amount of regularization with both tending to 0 as $n \to \infty$. Further discussion of tuning parameters is given in Section 2.4.

To construct CIs for $\boldsymbol{\theta}_0$, we may rely on asymptotic results similar to those suggested in Fan and Li (2002) and Zhang and Lu (2007), or resampling methods such as the bootstrap. However, the asymptotic-based approach tends to underestimate the variability as shown in, for example, Minnier *and others* (2011), as well as in the simulation results in Section 3.1. Thus, we turn to resampling to obtain more accurate assessment of the variability. First, we consider the commonly used wild bootstrap approach (Kosorok, 2007). We generate a vector of iid mean-1-variance-1 random variables, $\mathcal{V} = (\mathcal{V}_1, \ldots, \mathcal{V}_n)^{\mathrm{T}}$, independently of $\mathcal{O}$, and calculate

$$\hat{\boldsymbol{\theta}}^* = (1 + \lambda_2^*) \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ -\hat{\ell}_0^*(\boldsymbol{\theta}) + \lambda_2^* \sum_{j=p_D+1}^{p} \theta_j^2 + \lambda_1^* \sum_{j=p_D+1}^{p} \hat{w}_j^* |\theta_j| \right\},$$

where $\hat{w}_j^* = |\tilde{\theta}_{Rj}^*|^{-1}$, $\tilde{\boldsymbol{\theta}}_R^* = (1 + \lambda_2^*) \operatorname{argmin}_{\boldsymbol{\theta}} \{-\hat{\ell}_0^*(\boldsymbol{\theta}) + \lambda_2^* \sum_{j=p_D+1}^{p} \theta_j^2\}$, $\hat{\ell}_0^*(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \mathcal{V}_i \Delta_i [\boldsymbol{\theta}^{\mathrm{T}} \mathbf{W}_i - \log\{n \hat{\Pi}^{(0)*}(\boldsymbol{\theta}, X_i)\}]$, and $\hat{\Pi}^{(0)*}(\boldsymbol{\theta}, s) = n^{-1} \sum_{j=1}^{n} \mathcal{V}_j \mathbf{I}(X_j \geqslant s) \exp(\boldsymbol{\theta}^{\mathrm{T}} \mathbf{W}_j)$. In Appendix A of the Supplementary Material (available at *Biostatistics* online), we show that $\hat{\ell}_0(\boldsymbol{\theta})$ and $\hat{\ell}_0^*(\boldsymbol{\theta})$ are asymptotically equivalent to objective functions that are the sum of iid terms; thus, arguments similar to those given in Minnier *and others* (2011) show that $\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) \mid \mathcal{O} \overset{D}{\approx} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Thus, by producing B vectors $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(B)}$, we may find B iid estimators $\hat{\boldsymbol{\theta}}^{*(1)}, \ldots, \hat{\boldsymbol{\theta}}^{*(B)}$, and use the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}^{*(b)} - \hat{\boldsymbol{\theta}})$ to approximate the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, where B is some large number. In practice, we find good performance when $\mathcal{V}_i$ has finite support, such as $\mathcal{V}_i \sim 4 \cdot \mathrm{Beta}(\frac{1}{2}, \frac{3}{2})$. One may directly make inference about $\boldsymbol{\theta}$ based on $\{\hat{\boldsymbol{\theta}}^{*(b)}, b = 1, \ldots, B\}$; however, variance estimators from this approach tend to be overly conservative for $\{\theta_{0j}, j \in \mathcal{A}^c\}$ leading to imprecise interval estimation for survival functions. To produce valid inference for all coefficients as well as predicted survival, we instead propose the following two-stage ensemble voting approach.

2.2 *Ensemble voting*

In stage 1, we obtain $\hat{\boldsymbol{\theta}}$ and $\{\hat{\boldsymbol{\theta}}^{*(b)}, b = 1, \ldots, B\}$ as described in Section 2.1. Then we let the perturbed estimators vote, so that for $j = p_D + 1, \ldots, p$, if at least $\mathfrak{p}_j$ of $\{\hat{\theta}_j^{*(b)}, b = 1, \ldots, B\}$ are zero, both the point and interval estimates of θ_j are set to $\{0\}$ for some $\mathfrak{p}_j \in (0, 1)$. Details on the choice of $\mathfrak{p}_j$ are given in Section 2.4. Let $\hat{\mathcal{A}}_V = \{j : B^{-1} \sum_{b=1}^{B} \mathbf{I}(\hat{\theta}_j^{*(b)} = 0) \leqslant \mathfrak{p}_j\}$ be the active set based on voting. Obviously, $\{1, \ldots, p_D\} \subset \hat{\mathcal{A}}_V$ since coefficients for $\mathbf{D}$ are not penalized. In stage 2, we repeat the aENET regularized fitting and resampling using the restricted data $\{(X_i, \Delta_i, \mathbf{W}_{\hat{\mathcal{A}}_V i}^{\mathrm{T}})^{\mathrm{T}}, i = 1, \ldots, n\}$, where, for any $p \times 1$ vector $\mathbf{W}$ and any set $\mathcal{A} \subset \{1, \ldots, p\}$, $\mathbf{W}_{\mathcal{A}}$ denotes the subvector of $\mathbf{W}$ corresponding to $\mathcal{A}$. Let $\hat{\boldsymbol{\theta}}_{V, \hat{\mathcal{A}}_V}$ and $\{\hat{\boldsymbol{\theta}}_{V, \hat{\mathcal{A}}_V}^{*(b)}, b = 1, \ldots, B\}$ denote the corresponding estimates of the coefficients for $\mathbf{W}_{\hat{\mathcal{A}}_V}$ from the observed data and perturbations.

Let $\hat{\boldsymbol{\theta}}_V$ denote the final two-stage point estimator for $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}_V^{*(b)}$ denote the resampled counterpart of $\hat{\boldsymbol{\theta}}_V$ based on $\mathcal{V}^{(b)}$. Then the elements of $\hat{\boldsymbol{\theta}}_V$ and $\hat{\boldsymbol{\theta}}_V^{*(b)}$, are set to zero for $j \notin \hat{\mathcal{A}}_V$; and the subvectors of $\hat{\boldsymbol{\theta}}_V$ and $\hat{\boldsymbol{\theta}}_V^{*(b)}$ excluding these elements are, respectively, set to $\hat{\boldsymbol{\theta}}_{V, \hat{\mathcal{A}}_V}$ and $\hat{\boldsymbol{\theta}}_{V, \hat{\mathcal{A}}_V}^{*(b)}$. The variability in $\hat{\boldsymbol{\theta}}_V^{*(1)}, \ldots, \hat{\boldsymbol{\theta}}_V^{*(B)}$ now more closely matches the empirical variability of $\hat{\boldsymbol{\theta}}_V$, as demonstrated in the simulation studies. Asymptotic oracle properties of these ensemble-based estimators are established in Appendix A of the Supplementary Material (available at *Biostatistics* online).

2.3 *Survival functions and CIs*

To predict survival probabilities for a future patient with $\mathbf{W} = \mathbf{w}_{\text{new}}$, we may estimate $\Lambda_0(t)$ based on Breslow's estimator (Breslow, 1972), $\hat{\Lambda}_0(t; \hat{\boldsymbol{\theta}}) = \int_0^t \hat{\Pi}^{(0)}(\hat{\boldsymbol{\theta}}, s)^{-1} \, \mathrm{d}\bar{N}(s)$, where $\bar{N}(s) = n^{-1} \sum_{i=1}^n \Delta_i I[X_i \leqslant s]$. Subsequently, we estimate the survival function $S(t_0; \mathbf{w}_{\text{new}})$ as

$$\hat{S}(t_0; \mathbf{w}_{\text{new}}) = g\{\log \hat{\Lambda}_0(t; \hat{\boldsymbol{\theta}}) + \hat{\boldsymbol{\theta}}^{\mathrm{T}} \mathbf{w}_{\text{new}}\}. \tag{2.1}$$

To construct pointwise CIs for $S(t_0; \mathbf{w}_{\text{new}})$ for $t_0 \in [t_1, t_2] \subset (0, \tau)$, one may estimate the variances based on the asymptotic properties of $\hat{\boldsymbol{\theta}}$ and $\hat{\Lambda}_0(\cdot)$, where τ is a time satisfying $P(X > \tau) > 0$. However, such an explicit estimation approach may underestimate the variability as the case for $\hat{\boldsymbol{\theta}}$ and also is infeasible when the goal is to obtain simultaneous CIs.

We propose to employ the resampling method for both pointwise and simultaneous CI estimation. Specifically, for each set of $\mathcal{V}$, we first obtain the perturbed estimate $\hat{\boldsymbol{\theta}}^*$ and then calculate $\hat{S}^*(t_0; \mathbf{w}_{\text{new}}) = g\{\log \hat{\Lambda}_0^*(t_0; \hat{\boldsymbol{\theta}}^*) + \hat{\boldsymbol{\theta}}^{*\mathrm{T}} \mathbf{w}_{\text{new}}\}$, where $\hat{\Lambda}_0^*(t; \hat{\boldsymbol{\theta}}^*) = \int_0^t \hat{\Pi}^{(0)*}(\hat{\boldsymbol{\theta}}^*, s)^{-1} \, \mathrm{d}\bar{N}^*(s)$, and $\bar{N}^*(t) = n^{-1} \sum_{i=1}^n \mathcal{V}_i I(X_i \leqslant t) \Delta_i$. We demonstrate in Appendix B of the Supplementary Materials (available at *Biostatistics* online) that $\sqrt{n}\{\hat{S}^*(t_0; \mathbf{w}_{\text{new}}) - \hat{S}(t_0; \mathbf{w}_{\text{new}})\} \mid \mathcal{O}$ and $\sqrt{n}\{\hat{S}(t_0; \mathbf{w}_{\text{new}}) - S(t_0; \mathbf{w}_{\text{new}})\}$ converge weakly to the same limiting zero-mean Gaussian process. Thus, we may use the observed realizations of $\hat{S}^*(t_0; \mathbf{w}_{\text{new}})$ to construct CIs for $S(t_0; \mathbf{w}_{\text{new}})$. The variance of $\hat{S}(t_0; \mathbf{w}_{\text{new}})$ may be estimated as $\hat{\sigma}_S(t_0; \mathbf{w}_{\text{new}})^2 = B^{-1} \sum_{b=1}^B \{\hat{S}^{*(b)}(t_0; \mathbf{w}_{\text{new}}) - \hat{S}(t_0; \mathbf{w}_{\text{new}})\}^2$. A 95% CI at t_0 may be calculated as $\hat{S}(t_0; \mathbf{w}_{\text{new}}) \pm 1.96 \hat{\sigma}_S(t_0; \mathbf{w}_{\text{new}})$. To construct simultaneous CIs, we follow the same strategy as in Lin *and others* (1994) based on the resampled realizations. A 95% simultaneous CIs over the range $[t_1, t_2]$ can be obtained as $\{\hat{S}(t; \mathbf{w}_{\text{new}}) \pm c_{95} \hat{\sigma}_S(t; \mathbf{w}_{\text{new}}), t \in [t_1, t_2]\}$, where c_{95} is the 95th percentile of the distribution of $[\sqrt{n} \sup_{t \in [t_1, t_2]} \{|\hat{S}^{*(b)}(t; \mathbf{w}_{\text{new}}) - \hat{S}(t; \mathbf{w}_{\text{new}})| / \hat{\sigma}_S(t; \mathbf{w}_{\text{new}})\}, b = 1, \ldots, B]$. In finite samples, coverage is improved if we calculate CIs on the logit scale.

2.4 *Implementation and tuning*

To obtain $\hat{\boldsymbol{\theta}}$ numerically, one may use the algorithm proposed in Wu (2012). Alternatively, one may use a quadratic approximation to the likelihood similar to those proposed in Wang and Leng (2007) and Zhang and Lu (2007) to convert to a penalized least squares problem. Specifically, for a given λ_2, let $\hat{\ell}_R'(\tilde{\boldsymbol{\theta}}; \lambda_2) = \partial \hat{\ell}_R(\boldsymbol{\theta}; \lambda_2)/\partial \boldsymbol{\theta}$ and $\hat{\ell}_R''(\tilde{\boldsymbol{\theta}}; \lambda_2) = \partial^2 \hat{\ell}_R(\boldsymbol{\theta}; \lambda_2)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}$. We take the Cholesky decomposition of $\hat{\ell}_R''(\tilde{\boldsymbol{\theta}}; \lambda_2) = \mathbb{X}^{\mathrm{T}} \mathbb{X}$, and define $\mathbb{Y} = (\mathbb{X}^{\mathrm{T}})^{-1}\{\hat{\ell}_R''(\tilde{\boldsymbol{\theta}}; \lambda_2)\tilde{\boldsymbol{\theta}} - \hat{\ell}_R'(\tilde{\boldsymbol{\theta}}; \lambda_2)\}$; we may check that $\frac{1}{2}(\mathbb{Y} - \mathbb{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbb{Y} - \mathbb{X}\boldsymbol{\theta}) \approx \hat{\ell}_R(\boldsymbol{\theta}; \lambda_2)$ up to constants. Thus, after a preliminary estimate $\tilde{\boldsymbol{\theta}} y_R$ is found, $\hat{\boldsymbol{\theta}} \approx (1 + \lambda_2) \operatorname{argmin}_{\boldsymbol{\theta}} \{-\frac{1}{2}(\mathbb{Y} - \mathbb{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbb{Y} - \mathbb{X}\boldsymbol{\theta}) + \lambda_1 \sum_{j=p_D+1}^p |\tilde{\theta}_{Rj}|^{-1} |\theta_j|\}$. We find that this approximation performs well in finite samples.

We need to select tuning parameters to ensure satisfactory performance of the regularized estimation and the resampling methods. To this end, we recommend employing weak L_2 regularization to avoid overshrinkage which can induce bias. For the L_1 regularization, we follow the same principles as suggested in Minnier *and others* (2011) and consider a modified BIC. Precisely, we select λ_2 to guarantee df^* degrees of freedom, using the implemented ridge option of `coxph` in R with $\mathrm{df}^* = 0.99 \times \min\{\sum_{i=1}^n \Delta_i, p\}$. We then select λ_1 by minimizing a modified BIC penalty, $\mathrm{BIC}(\lambda_1) = -2\hat{\ell}_0\{\hat{\boldsymbol{\theta}}(\lambda_1)\} + n^{0.1} \widehat{\mathrm{df}}(\lambda_1)$, where $\widehat{\mathrm{df}}(\lambda_1)$ is simply the number of non-zero elements of $\hat{\boldsymbol{\theta}}$ when λ_1 is used for tuning. We repeat tuning parameter selection for each perturbed estimate. When we recalculate the estimates after ensemble voting, $\hat{\boldsymbol{\theta}}_V, \hat{\boldsymbol{\theta}}_V^{(1)*}, \ldots, \hat{\boldsymbol{\theta}}_V^{(B)*}$, we use the same tuning parameters as used in the initial estimators $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^{(1)*}, \ldots, \hat{\boldsymbol{\theta}}^{(B)*}$.

To choose the proportions $\mathfrak{p}_j$ for determining whether W_j should be excluded, we propose here a data-driven approach that works well in practice, but note that any thresholds $\mathfrak{p}_j \in (0, 1)$ will yield the property

that $P\{B^{-1}\sum_{b=1}^{B}\mathbf{I}(\hat{\theta}_j^{*(b)}=0)\geqslant \mathfrak{p}_j \mid \mathcal{O}\}\to I(j\notin\mathcal{A})$ due to the oracle properties of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^*$. In simulation, the obvious choice $\mathfrak{p}_j=0.5$ for all j works relatively well; however, we find that an approach that is more tuned to the data yields improved performance in finite samples. Specifically, we use a permutation approach to estimate what the threshold would be under a global null, $H_0:\mathbf{Z}\perp T\mid\mathbf{D}$. Since $\mathbf{D}$ may be both associated with T and $\mathbf{Z}$, the standard permutation that breaks the link between $\mathbf{W}$ and (X,Δ) may not be ideal. To account for the correlation, we propose to first regress Z_j against $\mathbf{D}$ and obtain residuals $\tilde{Z}_j$ for $j=1,\ldots,p_Z$. Let $\tilde{\mathbf{Z}}=(\tilde{Z}_1,\ldots,\tilde{Z}_{p_Z})^{\mathrm{T}}$ be the new covariates. Then $\tilde{\mathbf{Z}}$ is uncorrelated with $\mathbf{D}$ and remains unrelated to T under H_0. Next, for each set of permuted data $\{(X_i,\Delta_i,\mathbf{D}_i^{\mathrm{T}},\tilde{\mathbf{Z}}_i^{\dagger^{\mathrm{T}}})^{\mathrm{T}},i=1,\ldots,n\}$, we fit the aENET regularized Cox model and perform resampling to obtain perturbed estimates of $\boldsymbol{\theta}_0$ under H_0, where $\{\tilde{\mathbf{Z}}_1^{\dagger},\ldots,\tilde{\mathbf{Z}}_n^{\dagger}\}$ represents permuted $\{\tilde{\mathbf{Z}}_1,\ldots,\tilde{\mathbf{Z}}_n\}$. If we perform M such permutations, and let $\mathfrak{p}_j^{(m)}$ denote the proportion of perturbed values that vote for each θ_j to be 0 from the mth permutation, we may calculate $\bar{\mathfrak{p}}_j=M^{-1}\sum_{m=1}^{M}\mathfrak{p}_j^{(m)}$. We then set $\mathfrak{p}_j=\min\{\max(0.05,\bar{\mathfrak{p}}_j),0.95\}$ to ensure that $\mathfrak{p}_j\in(0,1)$ for the ensemble voting. For ease of implementation, one may also simply choose a common threshold $\mathfrak{p}=p_Z^{-1}\sum_{j=1}^{p}\mathfrak{p}_j$ when the covariates are standardized, which is what we adopt in our numerical studies and seems to work well in practice.

3. Numerical studies

3.1 *Simulation studies*

To assess the performance of the proposed procedures, we generated $\mathbf{Z}$ from a multivariate normal distribution with mean 0 and compound symmetry structure with variance 1 and correlation ρ. We considered settings with $p_Z=10$, 20, and 30 covariates, and correlations $\rho=0$ and 0.5. For simplicity, we did not include any additional clinical covariates. The underlying signal was linear involving only the first five covariates; the structure of this signal was $h(\mathbf{z})=z_1+z_2+0.5(z_3+z_4+z_5)$. For each setting, we generated survival times under the Cox model $\lambda(t;\mathbf{z})=\lambda_0(t)\exp\{h(\mathbf{z})\}$, where $\lambda_0(t)$ is the hazard function from a Weibull($\lambda=1,k=3$). The censoring was generated from a uniform distribution with range chosen to produce ∼50% censoring. We considered small and moderate sample sizes ($n=200$ and $n=500$).

For prediction of survival time for future patients, we consider three individuals. One is the "baseline" individual ($\mathbf{W}^{(0)}$) who has all covariates equal to 0; for this individual, the estimate $\hat{\boldsymbol{\theta}}$ appears only in the estimation of the cumulative baseline hazard $\Lambda_0(\cdot)$. We also consider two individuals with non-trivial covariates, where $\hat{\boldsymbol{\theta}}$ appears twice in the survival function estimate (2.1). The individual $\mathbf{W}^{(1)}$, with covariate pattern $(0,0,2,2,2,0,\ldots,0)$, should emphasize difficulties in estimating the smaller signals. The individual $\mathbf{W}^{(2)}$, with covariate pattern $(-0.5,\ldots,-0.5)$, should emphasize overall difficulties in estimating both the non-zero and zero coefficients. We estimate the survival function and calculate CIs in the region $[t_1,t_2]$, where t_1 is defined to be the 10th percentile of X and t_2 is the 90th percentile of X. We present results on the CIs for the conditional survival function at $t_0=(t_1+t_2)/2$ as well as simultaneous CIs for $t\in[t_1,t_2]$.

We compare three methods for interval estimation: the bootstrap; our proposed approach using perturbation resampling and voting; and an approach mimicking the asymptotic method. For the asymptotic method, because formulas do not exist for CIs of the survival function, we mimicked the approach of the formula by applying aENET to the observed data, identifying which $\hat{\theta}_j$ are declared non-zero, and restricting to these covariates for estimation using resampling. Resampling methods use $B=2000$ resamples. Results presented are based on 2000 simulations.

In Figure 1, we present the biases and the empirical SEs of the two point estimators: the standard aENET estimator $\hat{\boldsymbol{\theta}}$ and our voting-based estimator $\hat{\boldsymbol{\theta}}_V$. For the coefficients $\theta_{0j}=0$, the absolute bias is

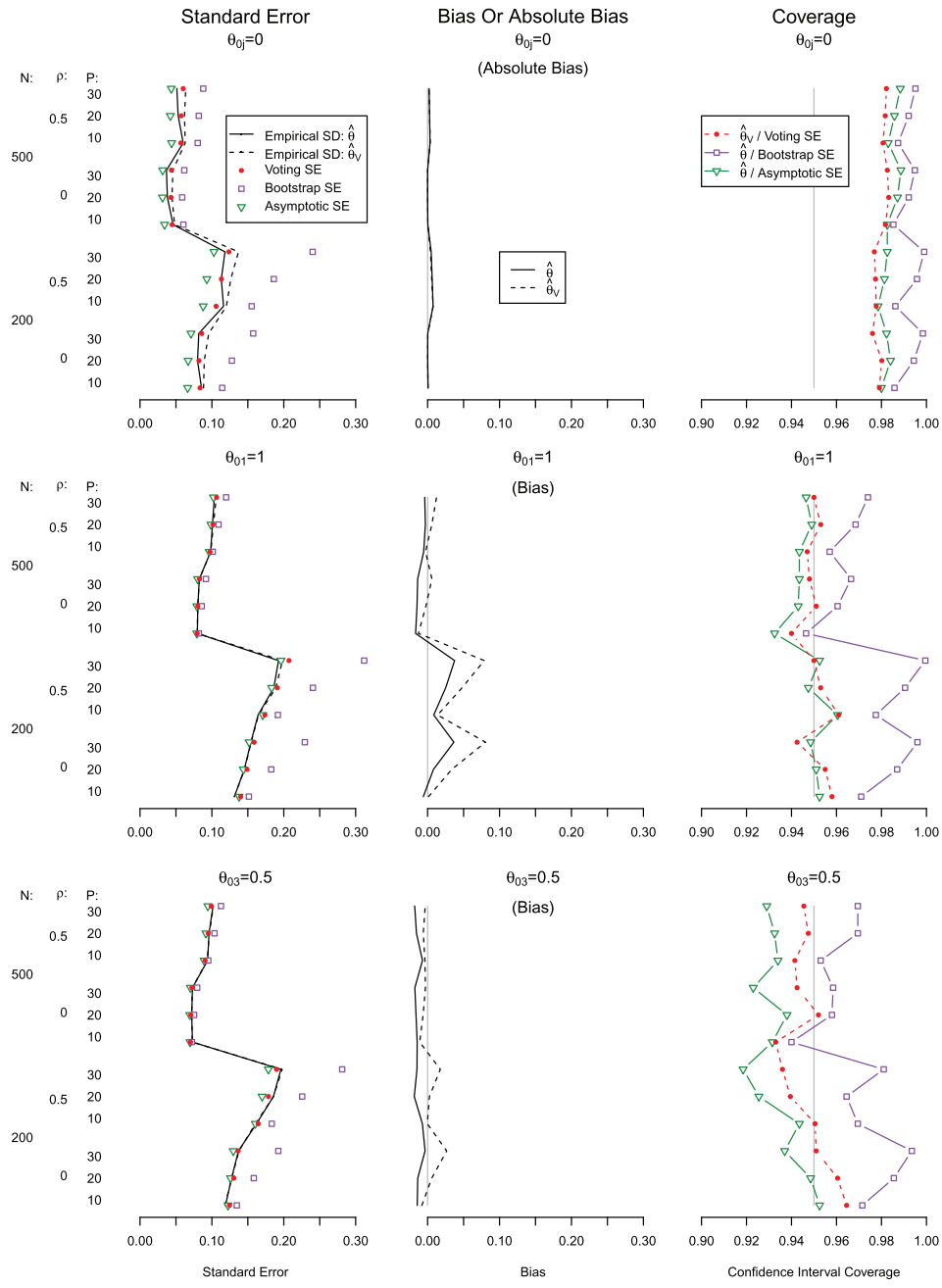

Fig. 1. Comparison of the SEs, bias, and 95% CI coverage of $\hat{\theta}_j$, for true model parameters $\boldsymbol{\theta}_0 = (1, 1, 0.5, 0.5, 0.5, 0, \ldots, 0)$. Shown are values when $\theta_{0j} = 0$ (with absolute bias displayed), as well as $\theta_{01} = 1$ and $\theta_{03} = 0.5$. Bias and empirical SEs are compared for the base aENET fit ($\hat{\boldsymbol{\theta}}$) and the aENET fit after the voting procedure ($\hat{\boldsymbol{\theta}}_V$); the variability for the base aENET fit may be estimated using either the bootstrap or the asymptotic method, while the variability for the voting procedure is estimated using the resampled coefficient estimators after voting.

displayed. The data-adaptive voting threshold $\mathfrak{p}$ varies between about 45% and 60%, depending on how "informed" the voters are—for example, when n is larger and p is smaller, a higher proportion of the voters successfully eliminate the true zeros under the global null, so a higher threshold $\mathfrak{p}$ may be used. Both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_V$ have negligible bias for zero and non-zero coefficients. For the non-zero signals, the bias is slightly upward for the strong signals ($\theta_{0j} = 1$) when $n = 200$ and $p = 20$ or 30. For the weaker signals ($\theta_{0j} = 0.5$), the aENET estimator has a downward bias as expected for any shrinkage estimator, but the voting-based estimator shifts the estimators slightly upward with slightly less bias for the weaker signals.

The empirical SEs for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_V$ are nearly identical for the non-trivial signals ($\theta_{0j} = 0.5$ or 1); when $\theta_{0j} = 0$, $\hat{\boldsymbol{\theta}}_V$ displays a slight increase in variability. This may be because the voting-based estimator actually tends to include covariates with true $\theta_{0j} = 0$ at a slightly higher rate than the standard aENET, although the frequencies of θ_j's being set to 0 only differ slightly between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_V$. Under no correlation, when $n = 200$, the percent of zero coefficients set to zero was (69%, 70%, 68%) for $\hat{\boldsymbol{\theta}}$ and (60%, 61%, 60%) for $\hat{\boldsymbol{\theta}}_V$ when $p_Z = (10, 20, 30)$. These frequencies get higher as expected when $n = 500$: (77%, 79%, 79%) for $\hat{\boldsymbol{\theta}}$ and (66%, 67%, 66%) for $\hat{\boldsymbol{\theta}}_V$. Under 0.5 correlation, when $n = 200$, the percent of zero coefficients set to zero was (70%, 70%, 67%) for $\hat{\boldsymbol{\theta}}$ and (69%, 62%, 60%) for $\hat{\boldsymbol{\theta}}_V$ when $p_Z = (10, 20, 30)$; when $n = 500$: (78%, 80%, 79%) for $\hat{\boldsymbol{\theta}}$ and (66%, 69%, 68%) for $\hat{\boldsymbol{\theta}}_V$. The slight over-selection of variables for $\hat{\boldsymbol{\theta}}_V$ is compensated with a small gain in retaining the true signals. Both methods always include the strong signals ($\theta_{0j} = 1$), but when $n = 200$, under no correlation $\hat{\boldsymbol{\theta}}$ misses moderate signals ($\theta_{0j} = 0.5$) (.05%, .06%, .03%) of the time, while $\hat{\boldsymbol{\theta}}_V$ misses them (.03%, 0%, .03%) of the time for $p = (10, 20, 30)$. Under 0.5 correlation, these rates are slightly higher, but with better success again for $\hat{\boldsymbol{\theta}}_V$: (1.1%, 1.9%, 2.0%) for $\hat{\boldsymbol{\theta}}$ and (0.7%, 1.4%, 1.6%) for $\hat{\boldsymbol{\theta}}_V$.

We have two methods to estimate $\text{var}\{\hat{\boldsymbol{\theta}}\}$ (asymptotic and bootstrap) and one to estimate $\text{var}\{\hat{\boldsymbol{\theta}}_V\}$ (our proposed resampling with voting method). When $\theta_{0j} = 0$, the variance calculated using asymptotics tends to fall below the empirical variance, while the bootstrap variance is typically higher. When $\theta_{0j} \neq 0$, the asymptotic method agrees with the empirical variance, while the bootstrap variance is still inflated when n is small. The ensemble-based voting method yields variance estimates that are more consistently in line with the empirical variance for all θ_{0j}.

Henceforth, we will compare CI coverage and refer to these by the error estimation method (asymptotic, bootstrap, and voting)—noting that asymptotic and bootstrap methods are centered at $\hat{\boldsymbol{\theta}}$ while voting-based methods are centered at $\hat{\boldsymbol{\theta}}_V$. The coverage for $\theta_{0j} = 0$ is high for all methods as expected based on the oracle properties. For $\theta_{01} = 1$, we see that the bootstrap method has substantial over-coverage due to overestimation of the variability, especially when $n = 200$, while the asymptotic and voting methods demonstrate near 95% coverage. For the moderate signal $\theta_{03} = 0.5$, the bootstrap intervals again tend to over-cover, and the asymptotic intervals exhibit some under-coverage when the number of covariates is larger. The ensemble voting method falls between these and maintains levels near 95%. In general, we find that our proposed voting method provides more precise estimation of the sampling variability compared to both the asymptotic based and bootstrap methods.

The coverage and width of the CIs for the t_0-year survival predictions are compared for the 3 individuals in Figure 2. The conditional survival probabilities at t_0 are approximately $(0.75, 0.01, 0.93)$ for the three individuals ($\mathbf{W}^{(0)}$, $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$). The bootstrap method tends to produce overly conservative CIs with coverage levels much higher than 95% and substantially broader widths. The asymptotic and voting-based CIs have very similar widths, but the voting-based coverage is typically higher. For $\mathbf{W}^{(0)}$, when all coefficients are 0, there is little difference between the asymptotic and voting methods. However, especially for $\mathbf{W}^{(1)}$, the asymptotic method can under-cover, while the voting-based CI has coverage near 95% across settings.

In Figure 3, we present results on the simultaneous CIs including their widths and empirical coverage levels. As with the pointwise intervals, we see that typically the simultaneous CIs based on bootstrap

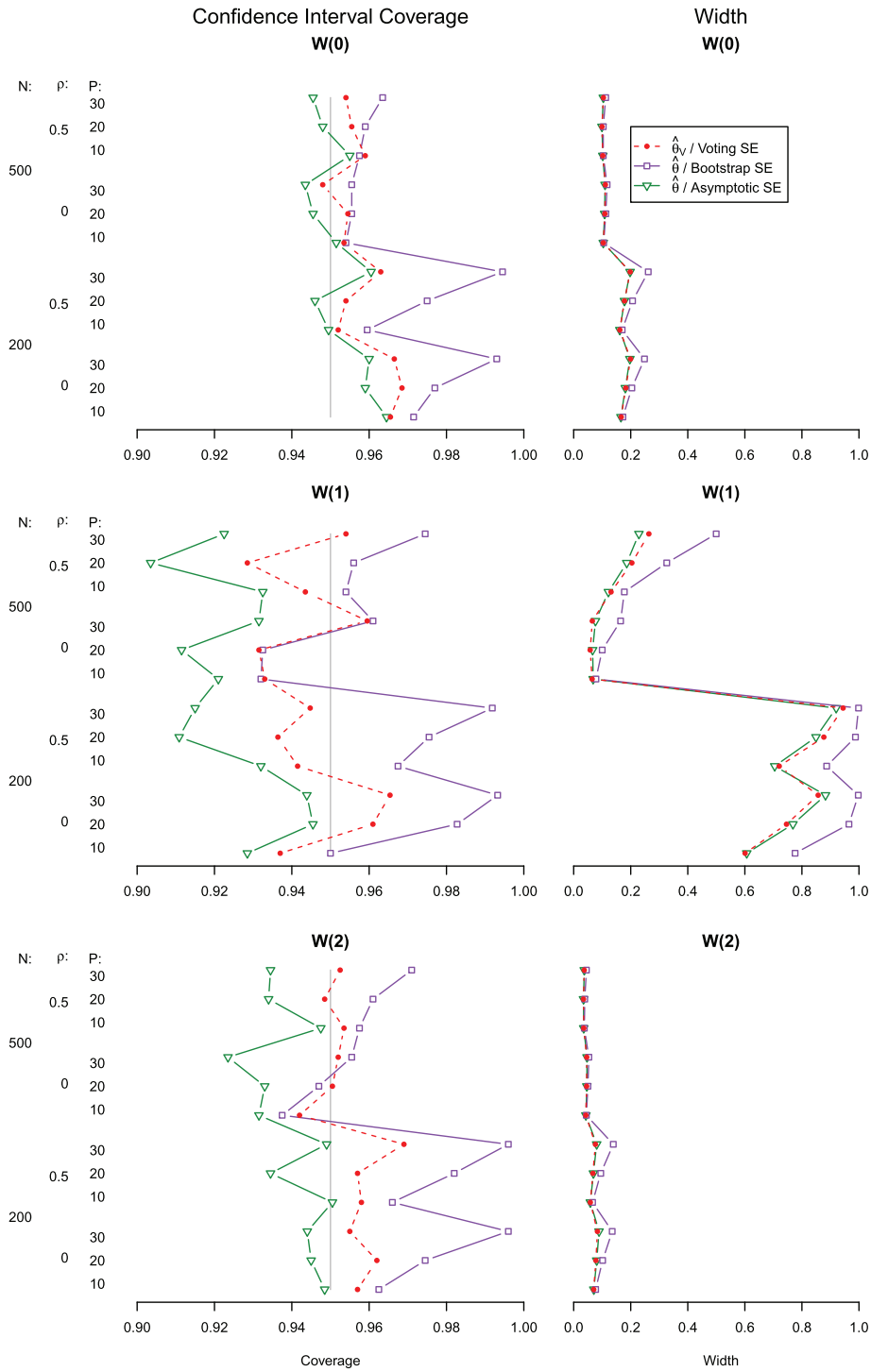

Fig. 2. Under the model with $\boldsymbol{\theta}_0 = (1, 1, 0.5, 0.5, 0.5, 0, \ldots, 0)$, CI coverage for t_0-year survival, and width, for three covariate levels: $\mathbf{W}^{(0)}$, with all covariates 0; $\mathbf{W}^{(1)} = (0, 0, 2, 2, 2, 0, \ldots, 0)$; and $\mathbf{W}^{(2)} = (-0.5, \ldots, -0.5)$.

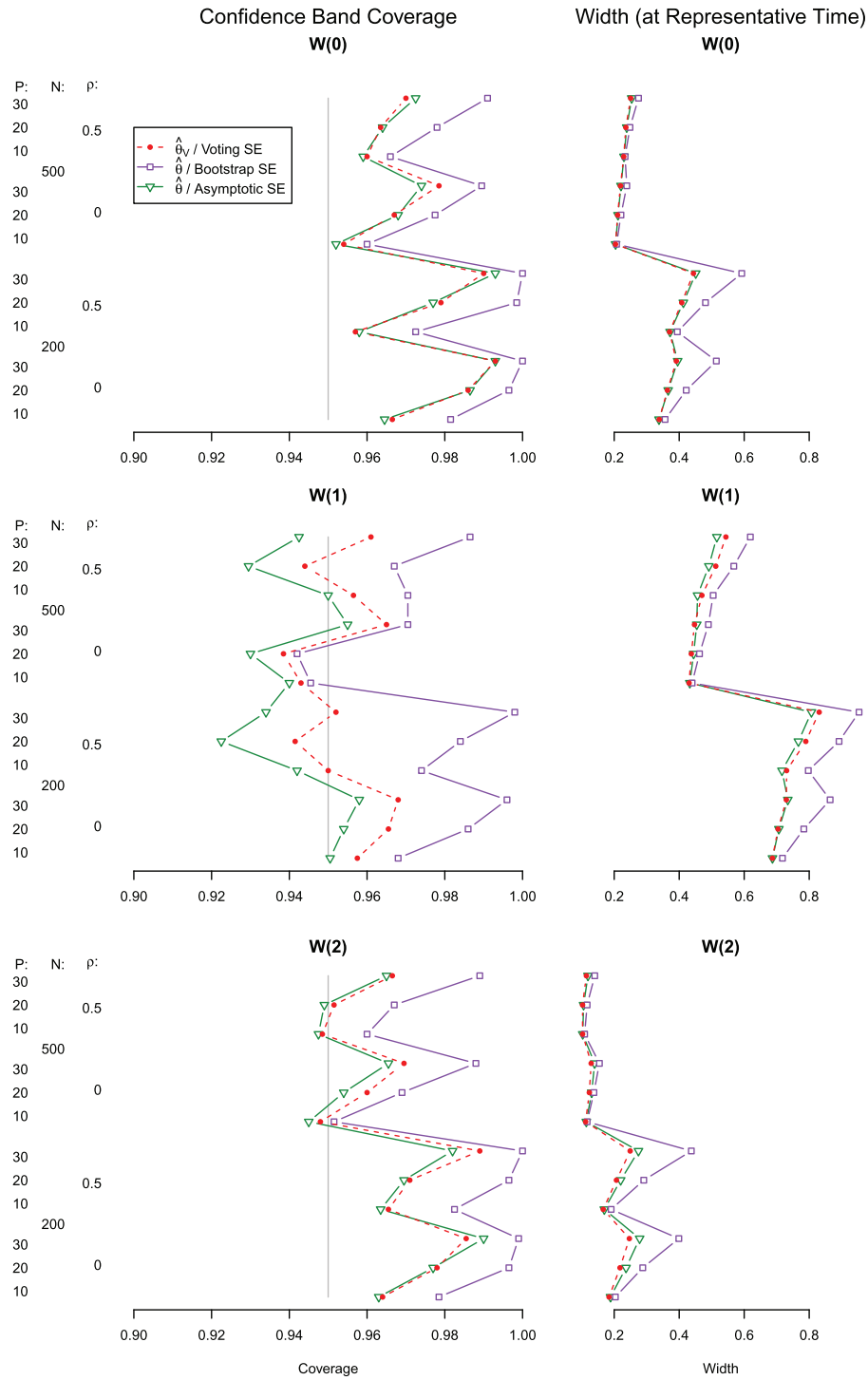

Fig. 3. Under the model with $\boldsymbol{\theta}_0 = (1, 1, 0.5, 0.5, 0.5, 0, \ldots, 0)$, simultaneous CI coverage for $\mathbf{W}^{(0)}$, with all covariates 0; $\mathbf{W}^{(1)} = (0, 0, 2, 2, 2, 0, \ldots, 0)$; and $\mathbf{W}^{(2)} = (-0.5, \ldots, -0.5)$. Also shown are simultaneous confidence widths at representative times.

over-cover and are wider, while the asymptotic and voting-based methods are narrower. When $n = 500$, coverage of the voting-based method is in general near 95%; when $n = 200$, the coverage tends to be over 95% for $\mathbf{W}^{(0)}$ and $\mathbf{W}^{(2)}$. For $\mathbf{W}^{(1)}$, we see where our voting method most improves over the asymptotic method because the ensemble voting provides more specific knowledge of which θ_{0j} should be set to 0.

The simulations presented above focused on settings where the true non-zero θ_{0j} were of large or moderate size, and thus almost always included in the models. We focus on this setting because precise interval estimation based on shrinkage estimators is feasible with moderate sample sizes. When some of the true signals are of order $O(n^{-1/2})$, adaptive lasso-type estimators are expected to yield significant bias for such weak signals and it becomes implausible to construct precise CIs as previously shown in Pötscher and Schneider (2009). To further examine the performance of our proposed procedures under such settings, we present additional simulation results in the Online Supplementary Materials (Web Appendix C, available at *Biostatistics* online) which follow the same structure as those above, but with $h(\mathbf{z}) = 1 \cdot z_1 + 0.8 \cdot z_2 + 0.6 \cdot z_3 + 0.4 \cdot z_4 + 0.2 \cdot z_5$, and focus in particular on the small signal $\theta_{05} = 0.2$. The probability of inclusion of the fifth variable in the support varies between 0.6 and 1.0, and we see that even the voting approach has empirical coverage levels significantly below the nominal level for θ_{05}, especially when $n = 200$. Once $n = 1000$, coverage returns to the nominal level for $\theta_{05} = 0.2$ (results not shown). For the survival functions, both asymptotic-based and bootstrap-based procedures tend to have difficulty in providing precise CIs under this setting, yielding either too low or too high of coverage levels. On the other hand, the proposed interval estimator for the survival functions based on ensemble and perturbation yields reasonable coverage levels despite the difficulty in making precise inference about the regression coefficients of weak signals. This further demonstrates the advantage of our proposed interval estimation procedures over existing methods based on asymptotic inference or bootstrap.

3.2 *Data example*

To illustrate our approach, we consider a breast cancer gene expression study previously reported in Wang *and others* (2005) consisting of 286 breast cancer subjects, 37% of whom experience breast cancer progression (107 events). We consider using the 62 genes belonging to the p53 signaling pathway (Subramanian *and others*, 2005) to predict breast cancer progression; this pathway is known to play an important role in breast cancer progression (Gasco *and others*, 2002). We build a model predicting survival, and compare the performance of the standard aENET-penalized Cox model with bootstrap-based CIs to our voting-based method of point and interval estimators. We are not presenting results based on the asymptotic formulas because our simulations suggested that they may not always be valid. We standardized each gene to have mean 0 and variance 1, and we included ER status as an unpenalized control covariate. Follow-up ranged between 2 months and 14.3 years; the range of observed deaths was between 2 months and 6.7 years. We provide predictions of survival between $t_1 = 9$ months and $t_2 = 5.1$ years.

Figure 4 shows the coefficient estimates with 95% CIs from the aENET using the bootstrap, as well the estimates and CIs from our perturbation with voting procedure. Our data-adaptive voting threshold $\mathfrak{p}_j$ was set to be 36%, so if more than 36% of the perturbations agreed that a coefficient should be 0, the covariate was excluded from the second stage of refitting. The genes in the model and the point estimates differ only slightly between the two methods, as we would expect based on the simulation studies: 28 genes are estimated to have non-zero effects in the initial aENET fit, while the voting-based estimate contains only 24 genes with non-zero effects. The real difference comes in the interval estimation. For example if we identify genes with nominal 95% CIs that exclude 0, we have only one significant gene (CHEK2) using the bootstrap, but five using the voting method (CHEK2, CCND3, CCNG1, CDKN2A, RRM2). CHEK2, a gene included on breast cancer hereditary panels, has been shown to interact with BRCA1 (Economopoulou *and others*, 2015). The other genes involved in cell cycle control are believed to relate to survival in numerous cancers including breast cancer; for example, CDKN2A was one of seven genes

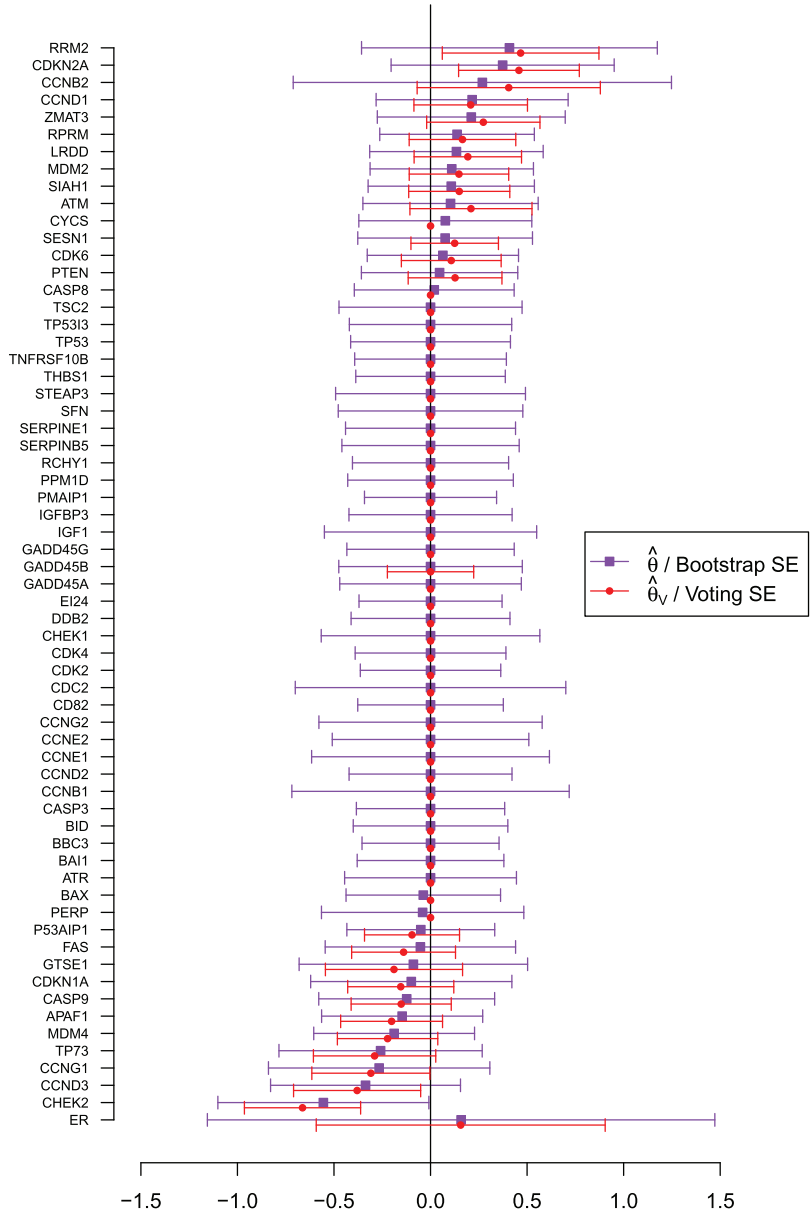

Fig. 4. In the breast cancer study, estimates of the (unpenalized) coefficient for ER status and the (penalized) coefficients for the variables in the p53 signaling pathway, each with 95% CIs, estimated using the aENET estimate with bootstrap CIs, and the voting-based method for both point estimation and interval estimation.

found to be useful for breast cancer progression prediction using a DNA methylation panel (Li *and others*, 2015). The point estimates of the log hazard ratios for these five genes according to the voting-based estimate are −0.66, −0.38, −0.31, 0.46, and 0.47, respectively. Thus, up-regularization of the first three genes is protective and of the last two genes is detrimental.

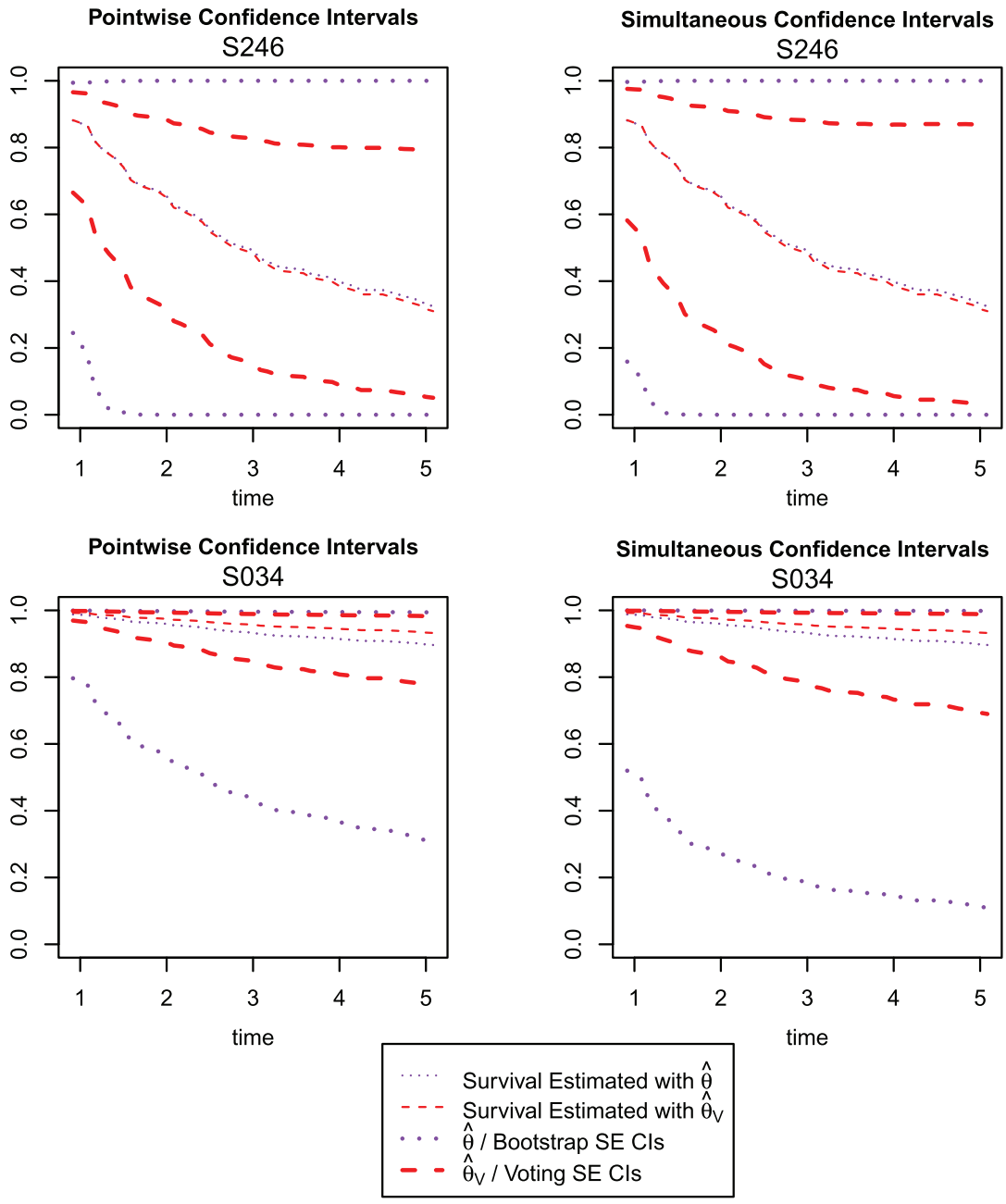

Fig. 5. Pointwise (left-hand column) and simultaneous (right-hand column) CIs for two individuals in the data set (top row: ID S246; bottom row: ID S034). The thin dotted line is the predicted survival from the aENET $\hat{\boldsymbol{\theta}}$; the thin dashed line is the predicted survival from voting-based estimate $\hat{\boldsymbol{\theta}}_V$. Thick dotted lines are the bootstrap-based confidence limits around the aENET predicted survival, and thick dashed lines are the voting-based confidence limits around the voting-based predicted survival.

To see how the difference in coefficient-level variability estimation impacts the variability estimation of downstream functions, we calculate the predicted survival function and associated CIs using the two methods for two individuals in the study, whose predicted 3-year survival rates are at the 10th and 90th percentiles; these are displayed in Figure 5. Individual S246 is at the 10th percentile; her predicted three-year survival probability is 0.49 using the aENET, with a 95% bootstrap CI of (0.00, 1.00)—a virtually meaningless interval. Her predicted three-year survival is very similar according to the voting-based estimate—0.48—but the 95% CI is narrower: (0.15, 0.83). The patient's breast tumor is ER positive, and the standardized gene expression values for the five genes (CHEK2, CCND3, CCNG1, CDKN2A, RRM2) are, respectively (-1.07, -1.69, -1.22, -0.25, -0.72). Individual S034 is at the 90th percentile; according to the aENET with bootstrap, her predicted three-year survival is 0.93, and a 95% CI is (0.44, 1.00). The voting-based predicted survival estimate is 0.96 with a much narrower 95% CI of (0.85, 0.99). The patient's breast tumor is ER positive, and the gene expression values for the fives genes listed above are (2.13, -0.15, -2.12, -1.33, -0.72). The pointwise and simultaneous CIs based on the voting procedure are dramatically narrower than the bootstrap-based limits, demonstrating the ability of our proposed method to give more precise but still accurate inferential information on the predicted survival.

4. Discussion

In this paper, we proposed an adaptation of a resampling approach to use with a penalization method for variable selection, in which we use an ensemble of resampled estimators $\hat{\boldsymbol{\theta}}^{*(1)}, \ldots, \hat{\boldsymbol{\theta}}^{*(B)}$ to better inform our knowledge of which θ_j are truly 0, with the goal of improving our estimation of the variability in $\hat{\boldsymbol{\theta}}$. We use the ensemble to vote out the unimportant covariates, and then by refitting the model in the data set and in resampled data sets, we produce estimators $\hat{\boldsymbol{\theta}}_V$ and $\hat{\boldsymbol{\theta}}_V^{*(1)}, \ldots, \hat{\boldsymbol{\theta}}_V^{*(B)}$ that can be used for valid inference. This not only improves precision of interval estimation for regression coefficients but also provides more precise interval estimation for the survival functions when compared with the standard bootstrap or asymptotic-based calculations. The voting-based perturbation approach tends to be the most robust across simulations, and maintains fairly good coverage levels with smaller interval width than the standard bootstrap. The compromise $\hat{\boldsymbol{\theta}}_V$ makes in variable selection enables us to reduce the downward bias for weak shrinkage and provide more accurate estimation of the sampling variability via resampling. In the context of risk prediction, prediction performance measures such as C-statistics (Uno *and others*, 2011) are often of interest for validating prediction models. Extending the proposed method to make precise inference about prediction accuracy measures warrants further research.

The actual mechanics of the voting based on resampling are similar to those proposed in other work. For example, Zhu and Fan (2011) perform stepwise selection on bootstrapped samples of the data, and select variables to be included in the final model based on the bootstrapped samples. Bach (2008) bootstraps the data and performs lasso on each bootstrapped sample, fitting a final model using unconstrained ordinary least squares on the variables that are non-zero in *every* bootstrap lasso fit. Meinshausen and Bühlmann (2010) subsample the data, perform lasso with a randomized weight, and then include variables that appear in some proportion of these fits. Our proposed method differs in some key details from these—we build around a variable selection method that has oracle properties in order to guarantee good asymptotic behavior and choose the voting threshold in a data-adaptive manner for good finite sample performance—but the general idea is similar. The main difference is that in other ensemble methods, the goal is typically improvement of variable selection and prediction; in contrast, we use the ensemble-derived knowledge to refit the model in both the original and the resampled data, in order to use the resampled data to more accurately assess error; to our knowledge, this has not been done previously. This allows us to achieve our goal of improving inference on potentially complicated functions of the parameter, such as the predicted survival.

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

Conflict of Interest: None declared.

Funding

J.A.S. was supported by the National Institutes of Health (NIH) grant T32 CA09001 and the A. David Mazzone Career Development Award. T.C. was supported by the NIH grants R01 GM079330, R01 HL089778, and U54 H6007963.

References

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In: *Proceedings of the 25th International Conference on Machine Learning*. New York: ACM, pp. 33–40.

Breslow, N. E. (1972). Contribution to the discussion of the paper by DR Cox. *Journal of the Royal Statistical Society, Series B* **34**(2), 216–217.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B* **34**, 187–220.

Economopoulou, P., Dimitriadis, G. and Psyrri, A. (2015). Beyond brca: new hereditary breast cancer susceptibility genes. *Cancer Treatment Reviews* **41**(1), 1–8.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**(1), 74–99.

Gasco, M., Shami, S. and Crook, T. (2002). The p53 pathway in breast cancer. *Breast Cancer Research* **4**(2), 70–76.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28**(5), 1356–1378.

Kosorok, M. R. (2007) *Introduction to Empirical Processes and Semiparametric Inference*. Berlin: Springer.

Li, Y., Melnikov, A. A., Levenson, V., Guerra, E., Simeone, P., Alberti, S. and Deng, Y. (2015). A seven-gene cpg-island methylation panel predicts breast cancer progression. *BMC Cancer* **15**(1), 417.

Lin, D. Y., Fleming, T. R. and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**(1), 73–81.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis* **52**(1), 374–393.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4), 417–473.

Minnier, J., Tian, L. and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association* **106**(496), 1371–1382.

Pötscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive lasso estimator. *Journal of Statistical Planning and Inference* **139**(8), 2775–2790.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. *and others* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**(43), 15545–15550.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**(4), 385–395.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. and Wei, L. J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**(10), 1105–1117.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-vanGelder, M. E., Yu, J. *and others* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365**(9460), 671–679.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association* **102**(479), 1039–1048.

Wu, Y. (2012). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Statistica Sinica* **22**, 27.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**(3), 691–703.

Zhu, M. and Fan, G. (2011). Variable selection by ensembles for the Cox model. *Journal of Statistical Computation and Simulation* **81**(12), 1983–1992.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* **37**(4), 1733.

[*Received March 23, 2015; revised December 17, 2015; accepted for publication March 23, 2016*]