Published in final edited form as:

Stat Biosci. 2017 December ; 9(2): 431–452. doi:10.1007/s12561-016-9170-z.

A LASSO Method to Identify Protein Signature Predicting Post-transplant Renal Graft Survival

Ling Zhou,

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA,
zholing@umich.edu

Lu Tang,

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USAW,
lutang@umich.edu

Angela T. Song,

College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI 48109, USA,
atsong@umich.edu

Diane M. Cibrik, and

Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA,
dcibrik@umich.edu

Peter X.-K. Song

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA,
pxsong@umich.edu

Abstract

Identifying novel biomarkers to predict renal graft survival is important in post-transplant clinical practice. Serum creatinine, currently the most popular surrogate biomarker, offers limited information of the underlying allograft profiles. It is known to perform unsatisfactorily to predict renal function. In this paper, we apply a LASSO machine-learning algorithm in the Cox proportional hazards model to identify promising proteins that are associated with the hazard of allograft loss after renal transplantation, motivated by a clinical pilot study that collected 47 patients receiving renal transplants at the University of Michigan Hospital. We assess the association of 17 proteins previously identified by Cibrik et al. [5] with allograft rejection in our regularized Cox regression analysis, where the LASSO variable selection method is applied to select important proteins that predict the hazard of allograft loss. We also develop a post-selection inference to further investigate the statistical significance of the proteins on the hazard of allograft loss, and conclude that two proteins KIM-1 and VEGF-R2 are important protein markers for risk prediction.

Keywords

L₁-penalty; Post-selection inference; Renal allograft loss; Regularized estimation; Survival analysis

1 Introduction

The survival rate of long term kidney allograft has not seen a dramatic improvement over the last decade or so, despite a significant decrease in acute rejection rates within one year of transplant attribute to immunosuppression [26]. It is incredibly important in nephrology to understand the biological mechanisms behind the occurrence of adverse events after renal transplants, particularly from the perspective of molecular biology. A vast amount of literature has been focused on the association between renal allograft dysfunction and a specific gene, genomic transcript, protein or their pattern [24,29,42,38,27,34,12,28,33,8,4,39,26,22,30,25]. Non-invasive biomarkers, such as serum protein signature, are much easier to measure and more cost-efficient, making them more appealing to nephrologists in disease diagnosis, treatment planning and decision making. However, very limited findings on specific gene, transcript or protein based on microarray data have been reported in the literature to this day. With few well-established biomarkers available, many studies in this field remain explorative by nature. The sample size of each study have been small due to limited financial sources for such research.

As far as proteomic biomarkers are concerned, it is believed that renal allograft dysfunction may be predicted by a group of proteins, the so-called protein signature, as opposed to specific individual proteins [5]. This hypothesis requires generating many individual proteins in an experiment, so there is a high chance of identifying a relevant protein group. As a result, detecting a promising protein signature presents a challenge in statistical analysis, as there are only sparse proteins from many candidates potentially associated with renal allograft profiles. Additionally, most of measured proteins are just noise. Beginning with 108 proteins, Cibrik et al. [5] successfully identified and validated 17 proteins that are able to differentiate rejecting renal transplants from stable transplants with high predictive power. In their analysis, the study endpoint was defined as a binary indicator of transplant categories, stable transplant (1) and allograft rejection (0). They used LASSO-type regularized logistic regression to perform variable selection among 108 predictors (i.e. proteins) with a sample of 47 subjects, the so-called large p small n scenario. Although the binary outcome is clinically relevant, a more important clinical endpoint is time to adverse event, which is primarily the interest in understanding long-term kidney allograft survival. Naturally, we further investigate which proteins, out of these 17 proteins previously identified by [5], can predict the hazard of allograft loss by analyzing survival data of time to allograft failure.

Many featured selection techniques for the linear regression model have been extended to the context of survival models, including the best-subset selection, stepwise selection, asymptotic procedures based on score tests, Wald test and other approximate chi-square testing procedures, bootstrap procedures [32], and Bayesian variable selection [11,18], among others. Variable selection via penalization has gained popularity since the mid-1990s. It uses some variable selection-capable penalty function to regularize the objective function while performing optimization. For the Cox model, Tibshirani [40] proposed a least absolute shrinkage and selection operator (LASSO), and Fan and Li [10] proposed the smoothly clipped absolute deviation (SCAD) method, which possesses an oracle property. Zhang and Lu [47] investigated the performance of adaptive LASSO [49] in Cox model, while Ma et al.

[23] and Wang et al. [45] studied group LASSO in Cox model for grouped predictors. These methods are further developed and applied in Cox regression with covariates missing at random [13], Cox models with varying coefficients [17], Cox models with ultra high-dimensional predictors [46], among many others.

The primary objective of this paper is to detect important proteins that enable the formation of protein signature that can predict the hazard of renal allo graft loss. In this case, we apply Cox proportional hazards models to analyze the survival data of the amount of time from kidney transplant to kidney failure, with predictors of 17 proteins as well as some confounding variables. Based on the assumption that protein signature is comprised of sparse signals, we hope to simultaneously remove unimportant proteins from the model. Traditional model selection procedures adopt the strategy of best-subset selection or some of the stepwise variants. However, such strategy is computationally prohibitive when the number of predictor is large. As demonstrated by [3], the subset selection method may be numerically unstable, so the resulting model may suffer from poor prediction accuracy. As one of the most popular variable selection methods, LASSO is able to overcome the computational hurdle of the subset selection approach. LASSO performs parameter estimation and variable selection simultaneously. Several fast algorithms have been developed for LASSO [9,36]. Due to both computational and statistical advantages, LASSO is our choice of method in this paper to establish a protein signature that is associated with the renal allograft survival.

To further refine the regularized estimation results, we generalize the high-dimensional post-selection inference proposed by van de Geer et al. [14] in the linear model to the Cox model. At a chosen Cox model under a selected tuning parameter, we provide both 95% confidence interval and p -value for the regression parameter. Note that this generalization is required for the post-selection inference because the LASSO estimation is obtained by penalizing the sparsity of protein signals, resulting in biased estimation.

This paper is organized as follows. We begin by a detailed description of the data collected from the pilot study in Section 2, where some summary statistics are provided. Section 3 presents the regularized Cox regression analysis with all 17 proteins included in one model, in which the LASSO regularization method estimates many regression coefficients exactly zero. Section 4 presents a new methodology of post-selection inference in the Cox model, which contains numerical results of simulations studies and data analysis. Section 5 provides some concluding remarks, including the presentation of detected proteins and their related functional annotation. Some technical details are included in the appendix.

2 Data Description

In a pilot study, Cibrik et al. [5] from the University of Michigan Hospital followed 47 patients who had received kidney transplants during 1990-2010 and collected information of adverse events associated with allograft loss. This group included those having either stable renal transplant or experiencing acute rejection of renal allograft. Stable renal transplant was defined as having normal and stable serum creatinine (Scr) levels ($\text{Scr} < 1.5 \text{ mg/dL}$) and without a clinical indication for a renal allograft biopsy (a Scr increase $\geq 25\%$ over the

baseline values). The Ban 07 classification was used to grade the rejection. A group of health controls recruited by the study will not be included in our survival analysis in this paper as they were not at-risk subjects. Patient's serum sample was analyzed at a hospital lab to measure protein expressions of 108 target markers using a customized antibody proteomics array.

Both demographic and clinical characteristics of these transplant patients are shown in Table 1. This patient sample had a median age of 51 years old, and was predominated by white Caucasian (85%, 39 white and 1 Asian) and by the first-time transplant (87%). On the other hand, the gender distribution appeared to be balanced (51% females and 49% males), so was the distribution of donor type (51% deceased donors versus 49% living donors).

To evaluate the hazard of post-transplant allograft loss, we calculated the time of duration from the day of transplant to the time of occurrence of a transplant failure event, kidney rejection or graft loss, whichever appeared first. Among 47 patients, 24 patients experienced either rejection or graft loss during the period of 2007 to 2014. For the two patients who had the date of graft loss but no recoded date of rejection, we simply defined the time to transplant failure based solely on the date of graft loss. For the remaining 23 patients who had neither rejection nor graft loss observed, their time to failure was treated as being censored. This leads to a censoring proportion of 48.9%.

As mentioned above, in this paper, we focus on 17 protein biomarkers that have been previously selected out of 108 proteins measured by an antibody microarray. For the detail concerning the proteomic lab technique and data cleaning, refer to [5]. Based on the binary outcome of rejection (yes or no), a logistic regression model was applied to screen the 108 proteins and a set of 17 proteins was found to be associated with the probability of allograft rejection. These 17 proteins (in an alphabetical order) include: epithelial calcium-dependent celladhesion protein (E-cadherin), epidermal growth factor (EGF), erythropoietin receptor (EPO-R), granulocyte-macrophage colony-stimulating factor (GM-CSF) growth regulated oncogene-alpha ($GRO\alpha$), interleukin 1 receptor 1 (IL-1-R1), interleukin 6 (IL-6), interleukin 12 p70 (IL-12p70), kidney injury molecule-1 (KIM-1), monocyte chemotactic protein-1 (MCP-1), macrophage migration inhibitory factor (MIF), macrophage inflammatory protein 3-alpha (MIP-3), bone sialoprotein 1 (Osteopontin), transforming growth factor β 1 (TGF- β 1), transforming growth factor β 2 (TGF- β 2), tumor necrosis factor receptor II (TNF-RII), and vascular endothelial growth factor receptor 2 (VEGF-R2). Their distributions are shown by the boxplot in Figure 1, where the order from the top to the bottom is determined by their p -values obtained via the univariate screening analysis, will be elaborated later, according to an ascending order of their p -values. The pairwise Spearman's correlations are reported in Table 2, and their statistical significance may be determined by the bootstrap resampling method (not shown in Table 2).

Our primary objective is to identify protein signature from this list of 17 markers that can predict the hazard of allograft loss. The use of 17 markers instead of the original 108 markers collected in the training dataset is due fully to the data availability; only these 17 markers are collected in the validation dataset for cost saving. In this paper we use the combined dataset of the training and validation data, in which only 17 proteins are available.

Yet these 17 markers have been repeatedly reported in the literature to be associated with multiple pathways of kidney disease. So, despite a smaller set of markers to be investigated, it is of great scientific importance to understand their roles in predicting the risk of allograft loss.

We begin with a simple univariate screening analysis that allows us to first rank these 17 proteins, from which we may gather the top single proteins predicting the hazard of allograft loss. To do so, we run the Cox proportional hazards model [6,7] with a single protein at one time, adjusted by confounding factors of gender, race, dialysis prior to transplant, age at transplant, medication, donor type and number of prior transplants. These proteins are ranked in an ascending order by their p -values, calculated by the likelihood ratio test, as shown in Table 3. To improve prediction power, it is appealing to include all 17 proteins in one Cox model, and in this way we are able to identify a set of proteins that jointly predict the hazard of allograft loss. The selection of important predictors can be carried out by the LASSO regularization approach. Then, we develop a post-selection inference to further examine the statistical significance of these 17 proteins in the regularized Cox regression analysis.

3 Regularized Cox Regression

Due to correlations among the proteins shown in Table 2, some top-ranked proteins may become insignificant when some strong predictors are present in the model. LASSO is useful to mitigate such a drawback, where the regularization estimation shrinks covariate effects to zero and allows stronger predictors to survive longer in the shrinkage process. For example, as reported later in this section, several promising proteins listed in Table 3, IL-1-R1, Osteopontin, GRO α and MIP-3 α , no longer play important roles in the multiple Cox model. Here we consider a Cox proportional hazards model:

$$\lambda(T_i|\mathbf{X}_i) = \lambda_0(T_i) \exp(\mathbf{X}_i^T \boldsymbol{\beta}), \quad (1)$$

where T_i is the duration from the date of transplant to the first occurrence of transplant failure event (i.e., either rejection or graft loss), $\mathbf{X}_i$ is a 24-element vector of predictors, including 17 proteins and 7 confounding variables, and is the corresponding vector of coefficients. Note that in this analysis, we do not regularize the estimation for the effects of the confounding variables. The conventional full model in this paper refers to the model given in (1) that contains all 24 covariates with no regularization. We fit such model and report the results in Table 4; see columns under names “CONV_EST” for the parameter estimates and “CONV_ p ” for the corresponding p -values. This full model is apparently not parsimonious as it contains many unimportant protein predictors.

To identify a sparse solution of protein signature predicting the hazard of graft loss, we apply the LASSO regularized estimation (R package glmnet2.0) in the Cox model. In fact, the LASSO estimator, adjusting for the confounding variables, is obtained by the following penalized optimization procedure:

$$\hat{\beta}(\lambda_n) = \arg \min_{\beta \in \mathbb{R}^p} \left(-\ell_n(\beta) + \lambda_n \sum_{j=1}^{17} |\beta_j| \right), \quad (2)$$

with the partial likelihood function,

$$\ell_n(\beta) = \frac{1}{n} \sum_{i=1}^n \delta_i \left[X_i^T \beta - \log \left\{ \sum_{j=1}^n \mathbf{I}(\mathcal{T}_j \geq \mathcal{T}_i) \exp(X_j^T \beta) \right\} \right]$$

where $\mathcal{T}_i = \min(T_i, C_i)$ is the observed event time, C_i is the censoring time defined as the duration from the date of transplant to the final date of sample collection, δ_i is the indicator representing failure time observed ($\delta_i = 1$) or not ($\delta_i = 0$). In addition, the LASSO L_1 -norm penalty is used in (2) where $\lambda_n \geq 0$ is the tuning parameter and $|a|$ is the absolute value of a . A large tuning parameter λ_n encourages sparsity in the estimation solution. That is, when $\lambda_n = 0$, the resulting estimation is the same as that given by the standard analysis with all 24 covariates included. When $\lambda_n = \infty$, the resulting estimates of $\beta_j, j = 1, \dots, 17$, would be all shrunk to zero, which is the case of no proteins being included in the Cox model. In reality, the tuning parameter is a value different from 0 or ∞ . In our analysis, the tuning parameter was chosen by a 30-fold cross-validation. This 30-fold cross-validation is chosen to ensure sufficient observations in training subsamples to achieve stable parameter estimation given such a low sample size. Through this cross-validation method, we obtained $\lambda = 0.096$. Based on this chosen tuning parameter, five proteins are identified as important predictors (i.e., with non-zero estimated effects) to the hazard of graft loss. They are, VEGF-R2, KIM-1, TNF-RII, E-Cadherin and EGF. The other 12 proteins in the list are not found as important predictors for the hazard of graft loss. See the results in the second column of Table 4 under name “LASSO_EST”.

The 17 solution paths of the regularized LASSO estimates are shown in Figure 2. From this figure, it is interesting to see that the coefficients of KIM-1, VEGF-R2 and TNF-RII have shown their resistance against being shrunk to zero, especially the first two. This means that they are really influential predictors. In contrast, the other two proteins E-cadherin and EGF have estimated effect sizes close to zero, showing much weaker prediction power than the other three.

4 Post-Selection Inference

From the above analysis results given by the LASSO Cox regression, a natural question arises: Are these five proteins with non-zero estimated effect sizes statistically significant? In this section, we develop a post-selection inference to provide an answer to this question.

4.1 Method

This section concerns a generalization of the post-selection inference developed by [14] from the linear model to the Cox model. This generalization provides the needed inference

tool to construct 95% confidence interval and to calculate p -values, and to perform hypothesis testing for the statistical significance of predictors after the model selection is settled. Unfortunately, the current large-sample theory of LASSO or other regularization methods does not provide applicable limiting distributions for inference; the bootstrap and other subsampling techniques have shown to be inappropriate for post-selection inference, either. Our goal is to establish pointwise statistical inference for each parameter β_j , $j = 1, \dots, p$, in Cox model (1). Following the derivation given by [14] for the linear model, we begin with the Karush-Kuhn-Tucker (KKT) condition of the LASSO in the Cox model, from which we give a bias-corrected estimator and consequentially establish its asymptotic normality.

For simplicity of notation, denote the LASSO estimator by $\hat{\beta} = \hat{\beta}(\lambda_n)$ in (2), and let β^0 be the true value of β . The KKT optimality conditions can be written as follows:

$$-\frac{1}{n} \sum_{i=1}^n \delta_i \left[X_i - \frac{\sum_{j=1}^n I(\mathcal{T}_j \geq \mathcal{T}_i) \exp(X_j^T \hat{\beta}) X_j}{\sum_{j=1}^n I(\mathcal{T}_j \geq \mathcal{T}_i) \exp(X_j^T \hat{\beta})} \right] + \lambda_n \hat{\kappa} = 0, \quad (3)$$

where the j -th element of the vector $\hat{\kappa}$ is given by

$$\hat{\kappa}_j \begin{cases} = \text{sign}(\hat{\beta}_j), & j=1, \dots, 17, \hat{\beta}_j \neq 0, \\ \in [-1, 1], & j=1, \dots, 17, \hat{\beta}_j = 0, \\ = 0, & j=18, \dots, 24. \end{cases} \quad (4)$$

Here, the vector $\hat{\kappa} \in \mathbb{R}^{24}$ is the subdifferential of the function $f(\beta) = \|\beta\|_1$ evaluated at $\beta = \hat{\beta}$. It is known that $\hat{\beta}$ is a LASSO solution in (2) if and only if $\hat{\beta}$ satisfies KKT conditions (3) and (4).

Adding the partial score function $s_n(\beta^0)$ (see Appendix A) on both sides of (3) and applying the first-order Taylor expansion, we obtain

$$I_n(\beta^0) (\hat{\beta} - \beta^0) + \lambda_n \hat{\kappa} \approx s_n(\beta^0), \quad (5)$$

where $I_n(\beta)$ is the observed Fisher information matrix defined in Appendix A. Substituting (3) into (5) leads to

$$\hat{\beta} - \beta^0 + I_n^{-1}(\beta^0) s_n(\hat{\beta}) \approx I_n^{-1}(\beta^0) s_n(\beta^0),$$

where $I_n(\beta)$ is assumed to be non-singular. At a given tuning parameter $\lambda_n < 0$, we define a bias-corrected estimator of β by $\hat{\beta}_{bc} = \hat{\beta} + I_n^{-1}(\hat{\beta}) s_n(\hat{\beta})$. Under some wild conditions, following the steps given in [14], we can show

$$\sqrt{n}(\hat{\beta}_{bc} - \beta_0) = W + o_p(1), \quad W|\mathbf{X} \sim \mathcal{N}_p\left(0, \{I(\beta^0)\}^{-1}\right),$$

where the Fisher information matrix $I(\beta)$ and the detailed proof are available in Appendix A. Since the true parameter β^0 is unknown, we plug in $\hat{\beta}$ into the observed Fisher information matrix to estimate the variance of $\hat{\beta}_{bc}$. An asymptotic pointwise $100(1 - \alpha)\%$ confidence interval for the j th element of β^0 , β_j^0 , is then constructed as follows,

$$\left[\hat{\beta}_{bc,j} - \sqrt{n}Z_{\alpha/2} \sqrt{I_n^{jj}(\hat{\beta})}, \quad \hat{\beta}_{bc,j} + \sqrt{n}Z_{\alpha/2} \sqrt{I_n^{jj}(\hat{\beta})} \right],$$

where $Z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution and $I_n^{jj}(\hat{\beta})$ denotes the j -th diagonal element of the inverse observed Fisher information matrix. It is interesting to note that, theoretically, the asymptotic variance of the bias-corrected estimator takes the same form as that of the conventional method with $\lambda = 0$. But in practice, we recommend plugging in the LASSO estimate instead of the conventional full-model estimates into the observed Fisher information for a better finite-sample performance, according to our experiences gained from both simulation studies and data analysis. However, as the sample size increases, the difference vanishes, as suggested by the theoretic results. See the details in the simulation studies below.

4.2 Simulation Study

In this section we examine the performance of the proposed method, denoted by LASSO_BC. In Tables 5 and 6, we report results of the comparison of LASSO_BC with standard LASSO (denoted by LASSO), adaptive LASSO (denoted by ALASSO), bias-corrected adaptive LASSO (denoted by ALASSO_BC), and the conventional Cox regression without penalty (denoted by CONV_Cox). The simulated data is generated from a Cox model $\lambda(t_i) = \exp(X_i^T \beta)$, $i=1, \dots, n$, where $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T \sim \mathcal{N}_p(0, \Sigma)$, with Σ being an exchangeable correlation matrix with correlation parameter γ ; $\beta = (\beta_1^T, \beta_2^T)^T$, $\beta_1 = (1.4, 1.8)^T$, which forms the signal set S_0 , $\beta_2 = \mathbf{0}$ is a zero vector of dimension $p-2$, namely the non-signal set S_0^c . The censoring indicator is generated from a binomial distribution with probability 0.3. Set the sample $n = 50, 100, 400$ and $p = 20$. All of the summary results are drawn from 500 rounds of simulations. The comparison takes places in the following domains:

- (i) Model selection sensitivity and model selection specificity. Here, model selection sensitivity is defined as the number of correctly selected variables divided by the true model size, and model selection specificity is defined as the number of correctly deleted variables divided by the the number of noise variables.
- (ii) Average bias for the signal set and that for the non-signal set, respectively,

$$\text{Avgbias}S_0 = \sum_{j \in S_0} |E\hat{\beta}_j - \beta_j|/2, \quad \text{Avgbias}S_0^c = (p - s_0)^{-1} \sum_{j \in S_0^c} |E\hat{\beta}_j - \beta_j|.$$

(iii) Average coverage probability of 95% confidence intervals CI_j for β_j

$$\text{Avgcov}S_0 = \sum_{j \in S_0} P[\beta_j \in CI_j]/2, \quad \text{Avgcov}S_0^c = (p - s_0)^{-1} \sum_{j \in S_0^c} P[0 \in CI_j].$$

(iv) Average length of 95% confidence intervals CI_j for β_j

$$\text{Avglength}S_0 = \sum_{j \in S_0} \text{length}(CI_j)/2, \quad \text{Avglength}S_0^c = (p - s_0)^{-1} \sum_{j \in S_0^c} \text{length}(CI_j),$$

All the results are shown in Table 5 and Table 6, corresponding to the respective cases of uncorrelated covariates with $\gamma = 0$ and correlated covariates with $\gamma = 0.8$.

When the sample size is small, a situation faced in our data analysis, the numerical

instability of calculating the inverse observed Fisher information $\hat{I}_n^{-1}(\hat{\beta})$ appeared to have potentially extra bias. To stabilize this estimation, we adopt an *ad hoc* ridge type method

based on a shrinkage estimator: $\tilde{I}_n^{-1}(\hat{\beta}) = \rho I_{p \times p} + (1 - \rho) I_n^{-1}(\hat{\beta})$, which is used to

substitute $I_n^{-1}(\hat{\beta})$ in the calculation of $\hat{\beta}_{bc}$, where $I_{p \times p}$ is the p -dimensional identity matrix. In the simulation, we compare the performance with $\rho = 0, 0.5, 1$, where obviously the original method is given by $\rho = 0$. From both Tables 5 and 6, we see that with $\rho = 0$ the coverage probabilities of LASSO_BC, as well as those of ALASSO_BC, tend to the nominal 95% level as the sample size increases, while for small sample sizes, the coverage probabilities of LASSO_BC or ALASSO_BC remain reasonably close to the nominal 95% level in the cases with $\rho = 0.5$ or 1, and the performance appears better in the cases of correlated covariates.

According to Table 5 and Table 6, all the methods considered have comparable high sensitivity even when the sample size is as low as 50. Three methods, CONV_Cox, LASSO_BC and ALASSO_BC, have higher specificity, compared to the standard LASSO and ALASSO methods. Lower specificity indicates a higher likelihood of overfitting, a known fact regarding the LASSO method. As far as the estimation accuracy concerns, the two bias-corrected estimators, LASSO_BC and ALASSO_BC, have shown smaller average biases over those selected variables, in comparison to the standard LASSO estimator or ALASSO estimator. Compared to the conventional full-model analysis (CONV_Cox), we find that the proposed LASSO_BC and ALASSO_BC have higher coverage probabilities in spite of shorter interval lengths. When the sample size is only 50, the coverage probability gets improved by 10% by using the proposed method, which is clear evidence in favor of the bias-corrected regularization approach. However, such difference becomes ignorable as the sample size increases, confirming the theoretic results. Finally, due to the fact that both

LASSO BC and ALASSO_BC are mostly comparable, we only apply LASSO_BC in our data analysis, taking the advantage of the computational simplicity of the LASSO method.

4.3 Data analysis

We apply the proposed post-selection inference tool to the analysis of the protein biomarker data. Both parameter estimates and p -values are listed in the columns of Table 4 under names LASSO_BC_EST and LASSO_BC_p, with the ridge shrinkage coefficient $\rho = 0, 1, 0.5$, respectively. Both the conventional full-model based inference and the proposed post-selection inference method declare KIM-1 and VEGF-R2 as significant predictors for the hazard of transplant failure. This leads to the protein signature identified from the analysis. Following [15], we perform a test to check the proportional hazards assumption for the final model including the two significant predictor, KIM-1 and VEGF-R2, as well as the confounding factors. The resulting p -value obtained is 0.1426, by R function `cox.zph` in package `survival`; this indicates the proportionality assumption is appropriate.

Utilizing these two significant predictors, we estimate the one-year allograft survival and the reduced proportion of the five-year survival relative to that of the one-year survival, using the final Cox model. Overall, we find that the one-year survival probability decreases as KIM-1 and/or VEGF-R2 increases. Specifically, the one-year survival would reduce to 0.90 and to 0.65 at, respectively, the baseline median values and the third quantiles of KIM-1 and VEGF-R2. In addition, the five-year survival would reduce nearly to 20% and 55% of the one-year survival at the baseline median values of KIM-1 and VEGF-R2 and at their third quantiles, respectively.

5 Concluding Remarks

Using proteins generated from a microarray, our study identifies a protein signature (i.e. a combination of KIM-1 and VEGF-R2) that has shown a strong association with the renal allograft survival for kidney transplant patients. The two identified significant protein biomarkers, KIM-1 and VEGF-R2, from our analysis have also shown great relevance to kidney disease in the literature. Previous studies have detected the expression of KIM-1 in the proximal tubule due to a toxin or pathophysiological state that results in dedifferentiation of the epithelium [19,43,48], which serves as early signal of the epithelial cell response to injury [1]. KIM-1 is also highly expressed in patients with a certain type of renal cell carcinoma [16]. In addition, expression of KIM-1 has been found in individuals with chronic renal failure in their proximal tubules, but it is not detected in the normal kidney [41,44,2]. VEGF-R2 is one of the receptors of vascular endothelial growth factor A (VEGF-A), which plays an important role in renal disease. Despite many studies that reported VEGF-A to be beneficial in several types of renal injury [37,20,21,35], stimulation of VEGF-R2, however, may be associated with renal injury as it is found to increase in damaged tubular epithelial cells and caused angiogenesis in the kidney [31].

Being two separate biomarkers, KIM-1 and VEGF-R2 are found to have individual significant positive associations with the hazard of renal failure. Clinical variables, such as race and dialysis status, are included as important confounding variables in the conventional Cox regression analysis. In the analysis based on the conventional full model, a

simultaneous inference for the 24 predictors based only on 47 observations may be unstable and prone to error. A more parsimonious model is appealing in the data analysis, which is achieved by utilizing the regularized LASSO method. This variable selection method offers us useful insights on which variables affect the hazard of renal allograft function loss. The LASSO regularized Cox regression enables the selection of a Cox model with the smallest cross-validation error. However, directly calling statistical significance on those non-zero estimated coefficients is inadequate. The post-selection inference method developed for the LASSO Cox regression in Section 4 gives a theoretically justified approach to reporting statistical significance, in which both confidence intervals and p -values are provided. It is interesting to note that although the standard LASSO method chooses 5 predictors according to non-zero estimated coefficients, only two of them remain statistically significant according to the proposed post-selection inference approach. To sum up, our post-selection inference method is useful to guarding against both false discoveries of “weak signals” and model overfitting resulted from the LASSO regularized Cox regression, an important step to further refine our analysis results and to report more reliable findings. Hence, the proposed post-selection inference for the Cox regression is a useful addition in the survival analysis.

There are several limitations to our study. The sample size in our study is small, which limits the ability to model interactions among the proteins. Also, when the sample size is small, the invocation of a sensitivity analysis over the ridge shrinkage coefficient takes some extra computational effort to examine the stability of discovery. The majority of patients in our sample are Caucasian, and only 15% of the patient sample are African American, who is known to have a high prevalence of kidney disease in the US. A methodological limitation is that some of the proteins may have non-linear effects on the hazard, which may not be captured by the linear effects analysis presented in this paper. In addition, we only focused on 17 proteins, instead of 108 proteins, due to limited data availability.

As for future work, non-linear and partial linear Cox model are useful to unveil non-linear relationships between proteins and hazard of renal allograft loss. An extension of the proposed inference method from the case of $p < n$ to a more challenging case of $p > n$ is worth exploring.

Acknowledgements

The authors are grateful to Associate Editor and two anonymous reviewers for their insightful comments and constructive suggestions that have led to an improvement of this paper. This research is supported in part by an NSF grant DMS#1513595 and an NIH grant R01-ES-024732.

Appendix A: Derivation of the asymptotic distribution of the bias-corrected estimator $\hat{\beta}^{bc}$

Denote the partial score function by $s_n(\beta) = \frac{1}{n} \sum_{i=1}^n \delta_i \left[X_i - \frac{\sum_{j=1}^n I(\mathcal{T}_j \geq \mathcal{T}_i) \exp(\mathbf{x}_j^T \beta) \mathbf{x}_j}{\sum_{j=1}^n I(\mathcal{T}_j \geq \mathcal{T}_i) \exp(\mathbf{x}_j^T \beta)} \right]$.

Denote $\hat{\beta} = \hat{\beta}(\lambda_n)$ given in (2), and let β^0 be the true value of β . It is known that the regularized estimator of β in (2) satisfies KKT condition (3), namely,

$$\lambda_n \hat{\kappa} = \frac{1}{n} \sum_{i=1}^n \delta_i \left[\mathbf{X}_i - \frac{\sum_{j=1}^n I(\mathcal{T}_j \geq \mathcal{T}_i) \exp(\mathbf{X}_j^T \hat{\beta}) \mathbf{X}_j}{\sum_{j=1}^n I(\mathcal{T}_j \geq \mathcal{T}_i) \exp(\mathbf{X}_j^T \hat{\beta})} \right]. \quad (6)$$

Adding $s_n(\beta^0)$ on both sides of (3) we obtain

$$\frac{1}{n} \sum_{i=1}^n \delta_i \left[\frac{S_{n2}(\mathcal{T}_i)}{S_{n0}(\mathcal{T}_i)} - \frac{S_{n1}^{\otimes 2}(\mathcal{T}_i)}{S_{n0}^2(\mathcal{T}_i)} \right] (\hat{\beta} - \beta^0) + \lambda_n \hat{\kappa} + O_p(\|\hat{\beta} - \beta^0\|_2^2) = s_n(\beta^0),$$

where $S_{nk}(y; \beta) = \sum_{j=1}^n I(\mathcal{T}_j \geq y) \exp(\mathbf{X}_j^T \beta) \mathbf{X}_j^{\otimes k}$ and $S_{nk}(y) = S_{nk}(y; \beta^0)$. Then the KKT conditions can be rewritten as follows,

$$\hat{\beta} - \beta^0 + \{I_n(\beta^0)\}^{-1} \lambda_n \hat{\kappa} = \{I_n(\beta^0)\}^{-1} s_n(\beta^0) + O_p(\|\hat{\beta} - \beta^0\|_2^2), \quad (7)$$

where the observed Fisher information matrix $I_n(\beta) = \frac{1}{n} \sum_{i=1}^n \delta_i \left[\frac{S_{n2}(\mathcal{T}_i; \beta)}{S_{n0}(\mathcal{T}_i; \beta)} - \frac{S_{n1}^{\otimes 2}(\mathcal{T}_i; \beta)}{S_{n0}^2(\mathcal{T}_i; \beta)} \right]$, which is assumed to be non-singular. For any vector $\mathbf{a}$, denote $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a} \mathbf{a}^T$.

Since the second term in (6) is asymptotically negligible, the following bias-corrected estimator is given by

$$\hat{\beta}_{bc} = \hat{\beta} + \{I_n(\beta^0)\}^{-1} \lambda_n \hat{\kappa} = \hat{\beta} + \{I_n(\beta^0)\}^{-1} s_n(\hat{\beta}), \quad (8)$$

where the second equality holds due to equation (6). Replacing with consistent estimates $I_n(\hat{\beta})$ and $s_n(\hat{\beta})$ for $I_n(\beta^0)$ and $s_n(\beta^0)$, respectively, we immediately obtain an asymptotic pivot. Denote the Fisher information matrix $\mathcal{I}(\beta) = E\{I_n(\beta)\}$. By central limit theory, the asymptotic distribution of $\hat{\beta}_{bc}$ follows from expression (8) under the same regularity conditions required for the consistency of the LASSO estimator.

References

1. Bonventre JV. Dedifferentiation and proliferation of surviving epithelial cells in acute renal failure. *Journal of the American Society of Nephrology*. 2003; 14(suppl 1):S55–S61. [PubMed: 12761240]
2. Bonventre JV. Kidney injury molecule-1 (kim-1): a urinary biomarker and much more. *Nephrology Dialysis Transplantation*. 2009; 24(11):3265–3268.
3. Breiman L, et al. Heuristics of instability and stabilization in model selection. *The annals of statistics*. 1996; 24(6):2350–2383.
4. Cheng O, Thuillier R, Sampson E, Schultz G, Ruiz P, Zhang X, Yuen PS, Mannon RB. Connective tissue growth factor is a biomarker and mediator of kidney allograft fibrosis. *American Journal of Transplantation*. 2006; 6(10):2292–2306. [PubMed: 16889607]

5. Cibrik DM, Warner RL, Kommareddi M, Song P, Luan FL, Johnson KJ. Identification of a protein signature in renal allograft rejection. *PROTEOMICS-Clinical Applications*. 2013; 7(11-12):839–849. [PubMed: 24323459]
6. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1972; 34(2):187–220.
7. Cox DR. Partial likelihood. *Biometrika*. 1975; 62(2):269–276.
8. Dadhania D, Snopkowski C, Ding R, Muthukumar T, Lee J, Bang H, Sharma VK, Seshan S, August P, Kapur S, Suthanthiran M. Validation of noninvasive diagnosis of bk virus nephropathy and identification of prognostic biomarkers. *Transplantation*. 2010; 90(2):189. [PubMed: 20526237]
9. Efron B, Hastie T, Johnstone I, Tibshirani R, et al. Least angle regression. *The Annals of statistics*. 2004; 32(2):407–499.
10. Fan J, Li R. Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*. 2002; 30(1):74–99.
11. Faraggi D, Simon R. Bayesian variable selection method for censored survival data. *Biometrics*. 1998; 54(4):1475–1485. [PubMed: 9883546]
12. Foxall PJD, Mellotte GJ, Bending MR, Lindon JC, Nicholson JK. Nmr spectroscopy as a novel approach to the monitoring of renal transplant function. *Kidney International*. 1993; 43:234–234. [PubMed: 8433564]
13. Garcia RI, Ibrahim JG, Zhu H. Variable selection for regression models with missing data. *Statistica Sinica*. 2010; 20(1):149. [PubMed: 20336190]
14. van de Geer S, Bühlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*. 2014; 42(3):1166–1202.
15. Grambsch PM, M TT. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994; 81(1):515–526.
16. Han WK, Bailly V, Abichandani R, Thadhani R, Bonventre JV. Kidney injury molecule-1 (kim-1): a novel biomarker for human renal proximal tubule injury. *Kidney International*. 2002; 62(1):237–244. [PubMed: 12081583]
17. Honda T, Härdle WK. Variable selection in cox regression models with varying coefficients. *Journal of Statistical Planning and Inference*. 2014; 148:67–81.
18. Ibrahim JG, Chen MH, MacEachern SN. Bayesian variable selection for proportional hazards models. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*. 1999; 27(4):701–717.
19. Ichimura T, Hung CC, Yang SA, Stevens JL, Bonventre JV. Kidney injury molecule-1: a tissue and urinary biomarker for nephrotoxicant-induced renal injury. *American Journal of Physiology-Renal Physiology*. 2004; 286(3):F552–F563. [PubMed: 14600030]
20. Kang DH, Hughes J, Mazzali M, Schreiner GF, Johnson RJ. Impaired angiogenesis in the remnant kidney model: Ii. vascular endothelial growth factor administration reduces renal fibrosis and stabilizes renal function. *Journal of the American Society of Nephrology*. 2001; 12(7):1448–1457. [PubMed: 11423573]
21. Kim YG, Suga S.i. Kang DH, Jefferson JA, Mazzali M, Gordon KL, Matsui K, Breiteneder-Gele S, Shankland SJ, Hughes J, Kerjaschki D, Schreiner GF, Johnson RJ. Vascular endothelial growth factor accelerates renal recovery in experimental thrombotic microangiopathy1. *Kidney International*. 2000; 58(6):2390–2399. [PubMed: 11115072]
22. Li B, Hartono C, Ding R, Sharma VK, Ramaswamy R, Qian B, Serur D, Mouradian J, Schwartz JE, Suthanthiran M. Noninvasive diagnosis of renal-allograft rejection by measurement of messenger rna for perforin and granzyme b in urine. *New England Journal of Medicine*. 2001; 344(13):947–954. [PubMed: 11274620]
23. Ma S, Song X, Huang J. Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*. 2007; 8(1):60. [PubMed: 17316436]
24. Magee CC, Denton MD, Womer KL, Khoury SJ, Sayegh MH. Assessment by flow cytometry of intracellular cytokine production in the peripheral blood cells of renal transplant recipients. *Clinical Transplantation*. 2004; 18(4):395–401. [PubMed: 15233816]
25. Mas V, Maluf D, Archer K, Yanek K, Mas L, King A, Gibney E, Massey D, Cotterell A, Fisher R, Posner M. Establishing the molecular pathways involved in chronic allograft nephropathy for

- testing new noninvasive diagnostic markers. *Transplantation*. 2007; 83(4):448–457. [PubMed: 17318078]
26. Muthukumar T, Dadhania D, Ding R, Snopkowski C, Naqvi R, Lee JB, Hartono C, Li B, Sharma VK, Seshan SV, Kapur S, Hancock WW, Schwartz JE, Suthanthiran M. Messenger rna for foxp3 in the urine of renal-allograft recipients. *New England Journal of Medicine*. 2005; 353(22):2342–2351. [PubMed: 16319383]
 27. Platt RE, Wu KS, Poole K, Newstead CG, Clark B. Soluble cd30 as a prognostic factor for outcome following renal transplantation. *Journal of Clinical Pathology*. 2009; 62(7):662–663. [PubMed: 19561239]
 28. Quintana LF, Solé-Gonzalez A, Kalko SG, Bañon-Maneus E, Solé M, Diekmann F, Gutierrez-Dalmau A, Abian J, Campistol JM. Urine proteomics to detect biomarkers for chronic allograft dysfunction. *Journal of the American Society of Nephrology*. 2009; 20(2):428–435. [PubMed: 19056874]
 29. Roberti I, Reisman L. Serial evaluation of cell surface markers for immune activation after acute renal allograft rejection by urine flow cytometry: Correlation with clinical outcome1. *Transplantation*. 2001; 71(9):1317–1320. [PubMed: 11397970]
 30. Sarwal M, Chua MS, Kambham N, Hsieh SC, Satterwhite T, Masek M, Salvatierra O Jr. Molecular heterogeneity in acute renal allograft rejection identified by dna microarray profiling. *New England Journal of Medicine*. 2003; 349(2):125–138. [PubMed: 12853585]
 31. Sato W, Tanabe K, Kosugi T, Hudkins K, Lanaspas MA, Zhang L, Campbell-Thompson M, Li Q, Long DA, Alpers CE, Nakagawa T. Selective stimulation of vegfr2 accelerates progressive renal disease. *The American Journal of Pathology*. 2011; 179(1):155–166. [PubMed: 21640973]
 32. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the cox regression model. *Statistics in Medicine*. 1992; 11(16):2093–2109. [PubMed: 12936711]
 33. Schaub S, Mayr M, Hönger G, Bestland J, Steiger J, Regeniter A, Mihatsch MJ, Wilkins JA, Rush D, Nickerson P. Detection of subclinical tubular injury after renal transplantation: comparison of urine protein analysis with allograft histopathology. *Transplantation*. 2007; 84(1):104–112. [PubMed: 17627245]
 34. Schaub S, Nickerson P, Rush D, Mayr M, Hess C, Golian M, Stefura W, Hay-Glass K. Urinary cxcl9 and cxcl10 levels correlate with the extent of subclinical tubulitis. *American Journal of Transplantation*. 2009; 9(6):1347–1353. [PubMed: 19459809]
 35. Shimizu A, Masuda Y, Mori T, Kitamura H, Ishizaki M, Sugisaki Y, Fukuda Y. Vascular endothelial growth factor165 resolves glomerular inflammation and accelerates glomerular capillary repair in rat anti-glomerular basement membrane glomerulonephritis. *Journal of the American Society of Nephrology*. 2004; 15(10):2655–2665. [PubMed: 15466270]
 36. Simon N, Friedman J, Hastie T, Tibshirani R, et al. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software*. 2011; 39(5):1–13.
 37. Suga SI, Kim YG, Joly A, Puchacz E, Kang DH, Jefferson JA, Abraham JA, Hughes J, Johnson RJ, Schreiner GF. Vascular endothelial growth factor (vegf121) protects rats from renal infarction in thrombotic microangiopathy. *Kidney International*. 2001; 60(4):1297–1308. [PubMed: 11576344]
 38. Sutherland SM, Li L, Sigdel TK, Wadia PP, Miklos DB, Butte AJ, Sarwal MM. Protein microarrays identify antibodies to protein kinase c ζ that are associated with a greater risk of allograft loss in pediatric renal transplant recipients. *Kidney International*. 2009; 76(12):1277–1283. [PubMed: 19812540]
 39. Tatapudi RR, Muthukumar T, Dadhania D, Ding R, Li B, Sharma VK, Lozada-Pastorio E, Seetharamu N, Hartono C, Serur D, Seshan SV, Kapur S, Hancock WW, Suthanthiran M. Noninvasive detection of renal allograft inflammation by measurements of mrna for ip-10 and cxcr3 in urine. *Kidney International*. 2004; 65(6):2390–2397. [PubMed: 15149352]
 40. Tibshirani R. The lasso method for variable selection in the cox model. *Statistics in Medicine*. 1997; 16(4):385–395. [PubMed: 9044528]
 41. van Timmeren MM, van den Heuvel MC, Bailly V, Bakker SJ, van Goor H, Stegeman CA. Tubular kidney injury molecule-1 (kim-1) in human renal disease. *The Journal of Pathology*. 2007; 212(2):209–217. [PubMed: 17471468]

42. Ting YT, Coates PT, Marti HP, Dunn AC, Parker RM, Pickering JW, Jack RW, Kemp RA, Walker RJ, McLellan AD. Urinary soluble hla-dr is a potential biomarker for acute renal transplant rejection. *Transplantation*. 2010; 89(9):1071–1078. [PubMed: 20098353]
43. Vaidya VS, Ramirez V, Ichimura T, Bobadilla NA, Bonventre JV. Urinary kidney injury molecule-1: a sensitive quantitative biomarker for early detection of kidney tubular injury. *American Journal of Physiology-Renal Physiology*. 2006; 290(2):F517–F529. [PubMed: 16174863]
44. Vaidya VS, Waikar SS, Ferguson MA, Collings FB, Sunderland K, Gioules C, Bradwin G, Matsouaka R, Betensky RA, Curhan GC, Bonventre JV. Urinary biomarkers for sensitive and specific detection of acute kidney injury in humans. *Clinical and Translational Science*. 2008; 1(3): 200–208. [PubMed: 19212447]
45. Wang S, Nan B, Zhou N, Zhu J. Hierarchically penalized cox regression with grouped variables. *Biometrika*. 2009; 96(2):307–322.
46. Wu TT, Wang S. Doubly regularized cox regression for high-dimensional survival data with group structures. *Statistics and Its Interface*. 2013; 6:175–186.
47. Zhang HH, Lu W. Adaptive lasso for cox's proportional hazards model. *Biometrika*. 2007; 94(3): 691–703.
48. Zhou Y, Vaidya VS, Brown RP, Zhang J, Rosenzweig BA, Thompson KL, Miller TJ, Bonventre JV, Goering PL. Comparison of kidney injury molecule-1 and other nephrotoxicity biomarkers in urine and kidney following acute exposure to gentamicin, mercury, and chromium. *Toxicological Sciences*. 2008; 101(1):159–170. [PubMed: 17934191]
49. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101(476):1418–1429.

Fig. 1.
Boxplots of the 17 standardized protein measurements.

Fig. 2. Solution paths of 17 proteins' parameter estimates from the LASSO Cox regression, adjusting for confounding variables.

Table 1

Summary statistics of demographic and clinical variables of 47 patients.

Variable		Descriptive statistics			
Gender	Male	23 (48.9%)	Female	24 (51.1%)	
Race	Black	7 (14.9%)	Others	40 (85.1%)	
Dialysis before tx	Yes	32 (68.1%)	No	15 (31.9%)	
Number of tx	1	41 (87.2%)	1+	6 (12.8%)	
Donor type	Living	23 (48.9%)	Deceased	24 (51.1%)	
Age at tx	Median	50.7	Range	[15.7, 72.4]	
Medication	Tacrolimus	39 (83.0%)	Cyclosporine	8 (17.0%)	

tx: transplant

Table 2

Pairwise Spearman's correlations among 17 proteins.

	VEGF	KIM	TNF	E_Cad	IL_1	EGF	Ost	GROa	MIP_3a	IL_12	MIF	TGF_B2	IL_6	TGF_B1	Epo_R	MCP_1	GM
VEGF_R2	1.00	0.47	0.52	-0.67	0.22	-0.56	0.52	-0.47	-0.60	0.33	0.28	-0.49	-0.46	-0.34	-0.38	-0.49	0.22
KIM_1	0.47	1.00	0.55	-0.65	0.78	-0.08	0.44	0.13	-0.00	0.77	0.67	0.14	0.04	-0.74	0.28	0.02	0.74
TNF_R2	0.52	0.55	1.00	-0.66	0.49	-0.29	0.43	-0.24	-0.36	0.41	0.37	-0.28	-0.31	-0.51	-0.11	-0.16	0.37
E_Cad	-0.67	-0.65	-0.66	1.00	-0.62	0.41	-0.55	0.29	0.47	-0.57	-0.60	0.25	0.35	0.69	0.08	0.25	-0.58
IL_1_R1	0.22	0.78	0.49	-0.62	1.00	0.08	0.26	0.26	0.20	0.91	0.85	0.40	0.16	-0.86	0.56	0.26	0.92
EGF	-0.56	-0.08	-0.29	0.41	0.08	1.00	-0.37	0.67	0.79	0.12	0.11	0.75	0.77	0.03	0.56	0.74	0.06
Ost	0.52	0.44	0.43	-0.55	0.26	-0.37	1.00	-0.56	-0.48	0.22	0.13	-0.37	-0.52	-0.15	-0.30	-0.50	0.25
GROa	-0.47	0.13	-0.24	0.29	0.26	0.67	-0.56	1.00	0.80	0.31	0.28	0.73	0.81	-0.24	0.78	0.74	0.28
MIP_3_a	-0.60	-0.00	-0.36	0.47	0.20	0.79	-0.48	0.80	1.00	0.18	0.10	0.78	0.81	-0.04	0.65	0.76	0.20
IL_12p70	0.33	0.77	0.41	-0.57	0.91	0.12	0.22	0.31	0.18	1.00	0.89	0.44	0.24	-0.84	0.59	0.25	0.90
MIF	0.28	0.67	0.37	-0.60	0.85	0.11	0.13	0.28	0.10	0.89	1.00	0.42	0.20	-0.84	0.58	0.28	0.86
TGF_B2	-0.49	0.14	-0.28	0.25	0.40	0.75	-0.37	0.73	0.78	0.44	0.42	1.00	0.81	-0.32	0.83	0.75	0.43
IL_6	-0.46	0.04	-0.31	0.35	0.16	0.77	-0.52	0.81	0.81	0.24	0.20	0.81	1.00	-0.15	0.70	0.76	0.18
TGF_B1	-0.34	-0.74	-0.51	0.69	-0.86	0.03	-0.15	-0.24	-0.04	-0.84	-0.84	-0.32	-0.15	1.00	-0.50	-0.24	-0.83
Epo_R	-0.38	0.28	-0.11	0.08	0.56	0.56	-0.30	0.78	0.65	0.59	0.58	0.83	0.70	-0.50	1.00	0.69	0.57
MCP_1	-0.49	0.02	-0.16	0.25	0.26	0.74	-0.50	0.74	0.76	0.25	0.28	0.75	0.76	-0.24	0.69	1.00	0.24
GM-CSF	0.22	0.74	0.37	-0.58	0.92	0.06	0.25	0.28	0.20	0.90	0.86	0.43	0.18	-0.83	0.57	0.24	1.00

Table 3

List of 17 proteins according to their p -values obtained from the univariate screening analysis based on the Cox regression, adjusted for clinical variables.

Protein	p -value
VEGF-R2	2.769×10^{-8}
KIM-1	9.045×10^{-7}
TNF-RII	4.823×10^{-6}
E-Cadherin	1.084×10^{-5}
IL-1-RI	1.456×10^{-4}
EGF	2.717×10^{-4}
Osteopontin	6.054×10^{-4}
GROα	1.223×10^{-3}
MIP-3α	2.341×10^{-3}
IL-12p70	9.550×10^{-3}
MIF	1.239×10^{-2}
TGF- β 2	1.377×10^{-2}
IL-6	2.929×10^{-2}
TGF- β 1	4.490×10^{-2}
Epo-R	1.099×10^{-1}
MCP-1	5.933×10^{-1}
GM-CSF	6.294×10^{-1}

Bold variables are significant after Bonferroni correction ($0.05/17 = 0.0029$).

Table 4

Parameter estimates and p values obtained by the conventional Cox regression and the LASSO regularized Cox regression. Covariates with p -values smaller than 0.05 are highlighted in bold.

Predictor	LASSO_EST	LASSO_BC_EST			LASSO_BC_p			CONV_EST	CONV_p
		$\rho = 0$	$\rho = 1$	$\rho = 0.5$	$\rho = 0$	$\rho = 1$	$\rho = 0.5$		
VEGF_R2	0.806	1.845	0.902	1.373	2.10×10^{-3}	1.33×10^{-1}	2.20×10^{-2}	4.229	4.73×10^{-4}
KIM_1	0.619	1.405	0.715	1.060	2.11×10^{-2}	2.41×10^{-1}	8.20×10^{-2}	4.020	2.34×10^{-3}
TNF_RII	0.342	0.113	0.437	0.275	8.03×10^{-1}	3.35×10^{-1}	5.44×10^{-1}	1.471	5.72×10^{-2}
E_Cadherin	-0.007	0.749	-0.102	0.323	3.57×10^{-1}	9.00×10^{-1}	6.91×10^{-1}	0.857	6.26×10^{-1}
IL_1_RI	0.000	0.255	0.098	0.177	7.15×10^{-1}	8.88×10^{-1}	8.01×10^{-1}	-1.368	2.48×10^{-1}
EGF	-0.025	0.008	-0.120	-0.056	9.91×10^{-1}	8.67×10^{-1}	9.38×10^{-1}	-0.971	6.34×10^{-1}
Osteopontin	0.000	-0.135	0.115	-0.010	7.86×10^{-1}	8.17×10^{-1}	9.84×10^{-1}	-1.253	2.04×10^{-1}
GRO_α	0.000	-0.416	-0.115	-0.265	6.85×10^{-1}	9.11×10^{-1}	7.96×10^{-1}	-2.564	2.79×10^{-1}
MIP_3_alpha	0.000	-1.125	-0.105	-0.615	1.11×10^{-1}	8.82×10^{-1}	3.84×10^{-1}	-0.773	5.68×10^{-1}
IL_12p70	0.000	-0.515	0.037	-0.239	5.43×10^{-1}	9.65×10^{-1}	7.78×10^{-1}	-1.828	2.33×10^{-1}
MIF	0.000	-0.181	0.036	-0.073	8.27×10^{-1}	9.66×10^{-1}	9.30×10^{-1}	0.739	5.29×10^{-1}
TGF_B2	0.000	1.518	-0.060	0.729	1.04×10^{-1}	9.49×10^{-1}	4.34×10^{-1}	2.418	2.92×10^{-1}
IL_6	0.000	-1.188	-0.128	-0.658	2.38×10^{-1}	8.99×10^{-1}	5.13×10^{-1}	-1.535	5.21×10^{-1}
TGF_β1	0.000	-1.280	-0.093	-0.687	9.72×10^{-2}	9.04×10^{-1}	3.74×10^{-1}	-3.289	9.38×10^{-2}
Epo_R	0.000	-0.202	-0.039	-0.121	8.11×10^{-1}	9.63×10^{-1}	8.87×10^{-1}	-0.533	6.42×10^{-1}
MCP_1	0.000	0.380	0.030	0.205	3.92×10^{-1}	9.47×10^{-1}	6.44×10^{-1}	0.313	3.90×10^{-1}
GM-CSF	0.000	-0.549	-0.026	-0.287	4.87×10^{-1}	9.74×10^{-1}	7.16×10^{-1}	-1.583	1.40×10^{-1}
Gender	0.145	0.117	0.145	0.131	7.71×10^{-1}	7.18×10^{-1}	7.44×10^{-1}	-0.090	8.63×10^{-1}
Race	-0.133	-0.471	-0.133	-0.302	1.06×10^{-1}	6.47×10^{-1}	3.00×10^{-1}	-1.101	2.38×10^{-2}
Dialysis prior tx	0.690	0.287	0.690	0.489	5.49×10^{-1}	1.50×10^{-1}	3.08×10^{-1}	1.813	2.19×10^{-2}
Number of tx	-0.224	-0.406	-0.224	-0.315	2.32×10^{-1}	5.09×10^{-1}	3.54×10^{-1}	-0.048	9.27×10^{-1}
Donor type	-0.018	-0.065	-0.018	-0.042	8.40×10^{-1}	9.55×10^{-1}	8.97×10^{-1}	-0.066	8.94×10^{-1}
Age at tx	-0.042	-0.116	-0.042	-0.079	7.46×10^{-1}	9.07×10^{-1}	8.26×10^{-1}	0.180	7.13×10^{-1}
Medicalim	0.036	0.192	0.036	0.114	6.94×10^{-1}	9.40×10^{-1}	8.15×10^{-1}	0.890	2.40×10^{-1}

Table 5

The sensitivity, specificity, average bias, average coverage probability and average length of 95% confidence interval of significant variables and non-significant variables for LASSO estimator (LASSO), ALASSO estimator (ALASSO), proposed post-selection estimators (LASSO_3C and ALASSO_3C), and the conventional full-model estimator (CONV_Cox) with no penalty, where the covariates are independent ($\gamma = 0$).

Measure	Method	$n = 50$	$n = 100$	$n = 400$
Sensitivity	LASSO	0.9990	1.0000	1.0000
	LASSO_BC($\rho = 0$)	0.9990	1.0000	1.0000
	LASSO_BC($\rho = 1$)	0.9710	1.0000	1.0000
	LASSO_BC($\rho = 0.5$)	0.9940	1.0000	1.0000
	ALASSO	1.0000	1.0000	1.0000
	ALASSO_BC($\rho = 0$)	0.9970	1.0000	1.0000
	ALASSO_BC($\rho = 1$)	0.9930	1.0000	1.0000
	ALASSO_BC($\rho = 0.5$)	0.9980	1.0000	1.0000
	CONV_Cox	0.9970	1.0000	1.0000
Specificity	LASSO	0.6922	0.6904	0.6034
	LASSO_BC($\rho = 0$)	0.8926	0.9283	0.9427
	LASSO_BC($\rho = 1$)	0.9959	0.9931	0.9824
	LASSO_BC($\rho = 0.5$)	0.9742	0.9746	0.9686
	ALASSO	0.8167	0.8760	0.9663
	ALASSO_BC($\rho = 0$)	0.8648	0.9186	0.9427
	ALASSO_BC($\rho = 1$)	0.9868	0.9860	0.9873
	ALASSO_BC($\rho = 0.5$)	0.9556	0.9651	0.9809
	CONV_Cox	0.8433	0.9162	0.9410
Avgbias S_0	LASSO	-0.4469	-0.3347	-0.1564
	LASSO_BC($\rho = 0$)	0.2864	0.1461	0.0451
	LASSO_BC($\rho = 1$)	-0.3597	-0.2694	-0.1260
	LASSO_BC($\rho = 0.5$)	-0.0366	-0.0616	-0.0404
	ALASSO	-0.0125	-0.0303	-0.0397
	ALASSO_BC($\rho = 0$)	0.4928	0.2188	0.0537
	ALASSO_BC($\rho = 1$)	0.0180	-0.0132	-0.0308
	ALASSO_BC($\rho = 0.5$)	0.2554	0.1028	0.0115
	CONV_Cox	0.9320	0.2891	0.0601
Avgbias S_0^c	LASSO	-0.0001	-0.0001	0.0001
	LASSO_BC($\rho = 0$)	-0.0026	-0.0002	-0.0001
	LASSO_BC($\rho = 1$)	-0.0005	-0.0000	-0.0000
	LASSO_BC($\rho = 0.5$)	-0.0015	-0.0001	-0.0001
	ALASSO	-0.0009	-0.0001	-0.0002
	ALASSO_BC($\rho = 0$)	-0.0032	-0.0004	-0.0001
	ALASSO_BC($\rho = 1$)	-0.0011	0.0001	-0.0002

Measure	Method	$n = 50$	$n = 100$	$n = 400$
Avgcov S_0	ALASSO_BC($\rho = 0.5$)	-0.0022	-0.0002	-0.0001
	CONV_Cox	-0.0049	-0.0006	-0.0001
	LASSO_BC($\rho = 0$)	0.8120	0.8550	0.9170
	LASSO_BC($\rho = 1$)	0.8100	0.7270	0.7320
	LASSO_BC($\rho = 0.5$)	0.9220	0.9100	0.9260
	ALASSO_BC($\rho = 0$)	0.7300	0.8220	0.9130
	ALASSO_BC($\rho = 1$)	0.9460	0.9310	0.9490
	ALASSO_BC($\rho = 0.5$)	0.8780	0.9140	0.9590
Avglength S_0	CONV_Cox	0.6280	0.7950	0.9120
	LASSO_BC	1.3519	0.8275	0.3849
	ALASSO_BC	1.5373	0.9001	0.3950
	CONV_Cox	2.1357	0.9935	0.4050
Avgcov S_0^c	LASSO_BC($\rho = 0$)	0.8926	0.9283	0.9427
	LASSO_BC($\rho = 1$)	0.9959	0.9931	0.9824
	LASSO_BC($\rho = 0.5$)	0.9742	0.9746	0.9686
	ALASSO_BC($\rho = 0$)	0.8648	0.9186	0.9427
	ALASSO_BC($\rho = 1$)	0.9868	0.9860	0.9873
	LASSO_BC($\rho = 0.5$)	0.9556	0.9651	0.9809
	CONV_Cox	0.8433	0.9162	0.9410
	LASSO_BC	1.0550	0.5858	0.2492
Avglength S_0^c	ALASSO_BC	1.1117	0.6009	0.2502
	CONV_Cox	1.2853	0.6172	0.2508

Table 6

The sensitivity, specificity, average bias, average coverage probability and average length of 95% confidence interval of significant variables and non-significant variables for LASSO estimator (LASSO), ALASSO estimator (ALASSO), proposed post-selection estimators (LASSO_3C and ALASSO_3C), and the conventional full-model estimator (CONV_Cox) with no penalty, where the correlation structure of covariates is compound symmetric with $\gamma = 0.8$.

Measure	Method	$n = 50$	$n = 100$	$n = 400$
Sensitivity	LASSO	0.9810	1.0000	1.0000
	LASSO_BC($\rho = 0$)	0.8000	0.9910	1.0000
	LASSO_BC($\rho = 1$)	0.4030	0.9480	1.0000
	LASSO_BC($\rho = 0.5$)	0.6730	0.9810	1.0000
	ALASSO	0.9360	1.0000	1.0000
	ALASSO_BC($\rho = 0$)	0.7990	0.9900	1.0000
	ALASSO_BC($\rho = 1$)	0.6020	0.9750	1.0000
	ALASSO_BC($\rho = 0.5$)	0.7280	0.9860	1.0000
	CONV_Cox	0.8210	0.9900	1.0000
Specificity	LASSO	0.7498	0.7359	0.7163
	LASSO_BC($\rho = 0$)	0.8880	0.9293	0.9447
	LASSO_BC($\rho = 1$)	0.9984	0.9972	0.9957
	LASSO_BC($\rho = 0.5$)	0.9844	0.9846	0.9828
	ALASSO	0.8161	0.8613	0.9686
	ALASSO_BC($\rho = 0$)	0.8700	0.9230	0.9443
	ALASSO_BC($\rho = 1$)	0.9902	0.9930	0.9950
	ALASSO_BC($\rho = 0.5$)	0.9650	0.9772	0.9908
	CONV_Cox	0.8470	0.9198	0.9442
Avgbias S_0	LASSO	-0.5240	-0.3638	-0.1855
	LASSO_BC($\rho = 0$)	0.3648	0.1572	0.0402
	LASSO_BC($\rho = 1$)	-0.4639	-0.3220	-0.1644
	LASSO_BC($\rho = 0.5$)	-0.0495	-0.0824	-0.0621
	ALASSO	-0.1789	-0.0700	-0.0599
	ALASSO_BC($\rho = 0$)	0.5102	0.1989	0.0440
	ALASSO_BC($\rho = 1$)	-0.1416	-0.0514	-0.0490
	ALASSO_BC($\rho = 0.5$)	0.1843	0.0737	-0.0025
	CONV_Cox	1.0358	0.2684	0.0515
Avgbias S_0^c	LASSO	0.0369	0.0259	0.0116
	LASSO_BC($\rho = 0$)	-0.0012	0.0034	0.0007
	LASSO_BC($\rho = 1$)	0.0599	0.0401	0.0177
	LASSO_BC($\rho = 0.5$)	0.0294	0.0218	0.0092
	ALASSO	0.0193	0.0074	0.0015
	ALASSO_BC($\rho = 0$)	-0.0023	0.0029	0.0006
	ALASSO_BC($\rho = 1$)	0.0365	0.0190	0.0100

Measure	Method	$n = 50$	$n = 100$	$n = 400$
Avgcov S_0	ALASSO_BC($\rho = 0.5$)	0.0171	0.0109	0.0053
	CONV_Cox	-0.0021	0.0033	0.0006
	LASSO_BC($\rho = 0$)	0.8730	0.9000	0.9410
	LASSO_BC($\rho = 1$)	0.9480	0.8940	0.8180
	LASSO_BC($\rho = 0.5$)	0.9650	0.9580	0.9520
	ALASSO_BC($\rho = 0$)	0.8370	0.8890	0.9400
	ALASSO_BC($\rho = 1$)	0.9350	0.9530	0.9800
	ALASSO_BC($\rho = 0.5$)	0.9180	0.9460	0.9730
Avglength S_0	CONV_Cox	0.7720	0.8800	0.9350
	LASSO_BC	2.5625	1.4415	0.6207
	ALASSO_BC	2.7007	1.4872	0.6253
	CONV_Cox	3.4971	1.5748	0.6337
Avgcov S_0^c	LASSO_BC($\rho = 0$)	0.8880	0.9293	0.9447
	LASSO_BC($\rho = 1$)	0.9984	0.9972	0.9957
	LASSO_BC($\rho = 0.5$)	0.9844	0.9846	0.9828
	ALASSO_BC($\rho = 0$)	0.8700	0.9230	0.9443
	ALASSO_BC($\rho = 1$)	0.9902	0.9930	0.9950
	ALASSO_BC($\rho = 0.5$)	0.9650	0.9772	0.9908
	CONV_Cox	0.8470	0.9198	0.9442
	LASSO_BC	2.3996	1.3072	0.5461
Avglength S_0^c	ALASSO_BC	2.4816	1.3263	0.5473
	CONV_Cox	2.9323	1.3656	0.5490