Cancer Informatics

ORIGINAL RESEARCH

Gene Selection using a High-Dimensional Regression Model with Microarrays in Cancer Prognostic Studies

Shuhei Kaneko, Akihiro Hirakawa and Chikuma Hamada

Department of Management Science, Graduate School of Engineering, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan. Corresponding author email: hirakawa@ms.kagu.tus.ac.jp

Abstract: Mining of gene expression data to identify genes associated with patient survival is an ongoing problem in cancer prognostic studies using microarrays in order to use such genes to achieve more accurate prognoses. The least absolute shrinkage and selection operator (lasso) is often used for gene selection and parameter estimation in high-dimensional microarray data. The lasso shrinks some of the coefficients to zero, and the amount of shrinkage is determined by the tuning parameter, often determined by cross validation. The model determined by this cross validation contains many false positives whose coefficients are actually zero. We propose a method for estimating the false positive rate (FPR) for lasso estimates in a high-dimensional Cox model. We performed a simulation study to examine the precision of the FPR estimate by the proposed method. We applied the proposed method to real data and illustrated the identification of false positive genes.

Keywords: cancer prognostic, false positive rate, gene selection, high-dimensional regression, microarray data, survival analysis

Cancer Informatics 2012:11 29–39

doi: 10.4137/CIN.S9048

This article is available from http://www.la-press.com.

Introduction

Establishing prognoses of clinical outcomes on the basis of microarray data is often performed in this decade.[1–4] In cancer research, not only the prediction of response to treatment but also the prediction of time to such events, eg, overall survival (OS) and relapse-free survival (RFS) are investigated.[5] To precisely predict such outcomes, we need to identify the genes that are highly correlated with them and are called the outcome-predictive genes. This is difficult, however, because the number of genes p in the high-dimensional microarray data exceeds the number of patients n. Several researchers have attempted to identify the outcome-predictive genes in the $n < p$ data settings by using traditional statistical methods, but the accuracy of the prediction based on the genes identified in this way is not very satisfactory. For example, van't Veer et al[3] and Van de Vijver et al[2] analyzed the gene expression profiles of 78 lymph node-negative breast cancer patients in order to establish gene signatures related to the risk of distant metastasis. Using a "three-step supervised classification method", they identified 70 genes that categorize patients into "good" and "bad" prognostic groups. Wang et al[4] also analyzed the gene expression profiles of 115 patients for the same purpose. They identified 76 genes by using the univariate Cox's proportional hazard regression analysis, which evaluates the relationship between the level of expression and the distant-metastasis-free survival for each gene. Notably, both studies had only 3 genes in common. Furthermore, the predictive performance based on both gene signatures drastically decreased when applied to other data sets.[6] Thus, the problem lies in the difficulty of precise identification of the outcome-predictive genes in high-dimensional data.

To address this difficulty, researchers have been emphasizing the penalized regression methods. Among them, the least absolute shrinkage and selection operator (lasso), which selects the outcome-predictive genes and simultaneously estimates the regression coefficients in the Cox regression model, is a typical penalized regression method.[7,8] This method shrinks all regression coefficients toward zero, and automatically sets many of them to exactly zero, depending on the amount of regularization employed. This can be useful, in particular, in high-dimensional data, and the prediction performance for microarray data have been widely studied by many researchers by using this method.[9,10] Several researchers showed that the lasso outperforms the simple variable selection methods such as the univariate Cox regression analysis,[9,11] with respect to the accuracy of prediction.

In the lasso, the amount of shrinkage varies, depending on the value of the tuning parameter, which is often determined by cross validation.[12] The number of genes selected as the outcome-predictive genes (ie, the genes included in the Cox model) generally decrease as the value of the tuning parameter increases. The optimal value of the tuning parameter that maximizes the prediction accuracy is determined; however, the set of genes identified using the optimal value contains the non-outcome-predictive genes (ie, false positive genes) in many cases.[10] Inclusion of such genes in the Cox model may yield an inaccurate prediction for the time-to-event outcome in patients. It is difficult to completely eliminate the false positive (FP) genes, even if we use the other penalized regression models. One idea to improve the identifiability of the outcome-predictive genes is to determine the FP genes, and subsequently, exclude them from the Cox model. To realize this, we developed a method for estimating the proportion of FP genes, ie, false positive rate (FPR), among the total identified genes. Specifically, the FPR is calculated using a mixture distribution based on the coefficients estimated by the lasso. We formulate the mixture distribution by considering the features of the lasso. By identifying the FP genes using the proposed method and excluding them from the Cox model, we are able to improve the prediction accuracy of the model. The accuracy of the FPR estimated by the proposed method is examined by simulation studies. We present the illustration of the proposed method using a well-known data set containing gene expressions from patients with diffuse large B-cell lymphoma (DLBCL) along with their survival time.[1]

We organize the remainder of this study as follows. In the Methods section, we introduce the Cox regression model, the lasso, and our proposed method for estimating the FPR. In the Simulation Studies section, we examine the accuracy of the estimated FPR through simulation studies with the situations observed in typical cancer prognostic studies with microarrays. In the

Application section, we illustrate the identification of FP genes by the proposed method by using the well-known DLBCL data. Finally, in the Discussion and Conclusion section, we discuss the characteristics of the proposed approach in further detail.

Methods

Cox proportional hazard model

Consider a sample of size n from which the relationship between the timing of an event and gene expression levels $x_1, \ldots, x_p$ of p genes need to be estimated. Due to censoring, for $i = 1, \ldots, n$, the ith datum in the patient is denoted by $(t_i, \delta_i, x_{i1}, \ldots, x_{ip})$, where δ_i is the censor indicator and t_i is the event time if $\delta_i = 1$ or censored time if $\delta_i = 0$, and $x_i = (x_{i1}, \ldots, x_{ip})^T$ is the vector of the gene expression levels of p genes for the ith patient. The Cox proportional hazard model is the most popular method to evaluate the relationship between gene expression and survival outcomes.[13] The hazard function of an event at time t for a patient i with the gene expression levels x_i is given by

$$h(t \mid x_i) = h_0(t) \exp(x_i^T \beta) \quad (1)$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a parameter vector and $h_0(t)$ is the baseline hazard, which is the hazard for the respective individual when all variable values are equal to zero. In the general setting with $n > p$, the coefficients are estimated by maximizing Cox's log partial likelihood as follows:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left[x_i^T \beta - \log \left\{ \sum_{r \in R(t_i)} \exp(x_i^T \beta) \right\} \right] \quad (2)$$

where $R(t_i)$ is the risk set that contains the patients whose survival time or censored time is at least t_i.

The lasso

Tibshirani[7,8] introduced a novel parameter estimating method that simultaneously executes parameter estimation and variable selection by adding the L1 norm to log likelihood function. The penalized likelihood function l_p of the lasso in the Cox's proportional hazard model is as follows:

$$l_p(\beta) = l(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \quad (3)$$

where λ is the tuning parameter, which determines the amount of shrinkage, and $l(\beta)$ is the Cox's log partial likelihood. The parameters are estimated by maximizing Equation (3). In this study, the parameters were estimated using the efficient gradient ascent algorithm.[14]

When performing the lasso, we need to determine the value of λ, which affects the lasso estimates. As the value of λ increases, the number of the selected genes monotonically decreases. The optimal value is often determined by the cross-validation log partial likelihood.[12] The K-fold cross-validated log partial likelihood is given by

$$CV(\lambda) = \sum_{k=1}^{K} \left\{ l(\hat{\beta}_{(-k)}) - l_{(-k)}(\hat{\beta}_{(-k)}) \right\} \quad (4)$$

where $l_{(-k)}(\hat{\beta})$ is the log partial likelihood when the kth fold is left out, and $\hat{\beta}_{(-k)}$ is the estimate of β obtained by the lasso when the kth fold is left out. The optimal tuning parameter λ is obtained by maximizing $CV(\lambda)$. The number of folds to execute the above-mentioned cross validation is often set to 5 (or 10), considering the computational feasibility.

Estimation of false positive rate (FPR)

In this section, we propose the method to estimate the FPR for a fixed value of λ determined by the cross validation by assuming a mixture distribution for the lasso estimates. The mixture distribution is developed on the basis of the following 2 features of the lasso: (i) the lasso selects at most n variables, because of the nature of the convex optimization problem when $n < p$,[15,16] and (ii) in the Bayesian framework, the lasso estimate is derived as the posterior mode under independent Laplace prior distribution as follows:

$$f_L\left(\beta_j; 0, \frac{1}{\lambda}\right) = \frac{\lambda}{2} \exp(-\lambda |\beta_j|) \quad (5)$$

where $f_L(y; a, b) = (2b)^{-1} \exp(-|y - a|/b)$ is the probability density function of Laplace distribution with location parameter a and scale parameter b.[7] On the basis of these features of the Lasso, the mixture

distribution is assumed to $\hat{\beta}_j$ for the fixed value of λ as follows:

$$f\left(\hat{\beta}_j;\pi_0,\pi_c,\tau,\mu_c,\sigma_c\right)=\frac{n}{p}\left\{\pi_0 f_L\left(\hat{\beta}_j;0,\tau^{-1}\right)+\sum_{c=1}^{c}\pi_c f_N\left(\hat{\beta}_j;\mu_c,\sigma_c^2\right)\right\}+\left(1-\frac{n}{p}\right)f_L\left(\hat{\beta}_j;0,\varepsilon\right) \quad (6)$$

where π_0 and π_c are mixed proportions $\left(\pi_0+\sum_{c=1}^{C}\pi_c=1\right)$; $f_N(\cdot;\mu_c,\sigma_c^2)$ is the probability density function of the normal distribution with mean μ_c ($\neq 0$) and variance σ_c^2 in component c; C is the number of component, which is determined on the basis of any model evaluation criteria; and ε is the constant value, which is boundlessly close to 0, eg, $\varepsilon = 10^{-8}$. The unknown parameters, π_0, π_c, τ, μ_c, and σ_c, are estimated by maximizing the log-likelihood function of Equation (6).

The mixture distribution defined by Equation (6) is formulated on the basis of the following concepts. Since the lasso selects at most n genes in the $n < p$ setting, the coefficients for at least $p - n$ genes are shrunken toward exactly zero; therefore, Equation (6) consists of 2 terms, ie, n/p term and $1 - n/p$ term. In the n/p term, the $C + 1$ component mixture distribution comprising the Laplace and normal distributions. Specifically, the Laplace distribution with location parameter 0 and scale parameter τ^{-1}, $f_L(\hat{\beta}_j;0,\tau^{-1})$, is assumed as the distribution of the non-outcome-predictive genes considering the above-mentioned feature (ii) of the lasso, while the C-component ($c = 1, \ldots, C$) normal distribution with mean μ_c ($\neq 0$) and variance σ_c^2 is assumed as the distribution of the outcome-predictive genes. It should be noted that normal distribution is a choice of convenience. Next, in the $1 - n/p$ term, the Laplace distribution with location parameter 0 and scale parameter ε is assumed as the distribution of the $p - n$ genes, considering the above-mentioned feature (i) of the lasso.

Using the estimated mixture distribution, we defined a FPR for a cut-off value ζ (>0) as follows: given the cut-off value ζ, the area under the Laplace distribution in the n/p term is the estimated proportion of FP genes, and can be written as follows:

$$\hat{P}_{\mathrm{FP}}=\hat{\pi}_0\left[\int_{-\infty}^{-\zeta}f_L\left(u;0,\hat{\tau}^{-1}\right)du+\int_{\zeta}^{+\infty}f_L(u;0,\hat{\tau}^{-1})du\right]=2\hat{\pi}_0\int_{\zeta}^{+\infty}f_L(u;0,\hat{\tau}^{-1})\,du \quad (7)$$

Next, the estimated proportion of true positive (TP) genes for the cut-off value ζ is given by the following:

$$\hat{P}_{\mathrm{TP}}=\sum_{c=1}^{C}\hat{\pi}_c\left[\int_{-\infty}^{-\zeta}f_N\left(u;\hat{\mu}_c,\hat{\sigma}_c^2\right)du+\int_{\zeta}^{+\infty}f_N(u;\hat{\mu}_c,\hat{\sigma}_c^2)\,du\right] \quad (8)$$

Using equation (7) and equation (8), we obtain the FPR estimator for the cut-off value ζ as follows:

$$\widehat{\mathrm{FPR}}(\zeta)=\frac{\hat{P}_{\mathrm{FP}}}{\hat{P}_{\mathrm{TP}}+\hat{P}_{\mathrm{FP}}} \quad (9)$$

Based on the cut-off value ζ used, the estimated proportions of FP and TP genes and the corresponding estimated FPR are found to vary. We determined a cut-off value based on the target FPR specified *in priori*. Specifically, by sequentially changing ζ, we determined the cut-off value that allowed the estimated FPR to be less than or equal to the target FPR. For example, if the target FPR was 0.05, we used the minimum cut-off value that would make the estimated FPR ≤ 0.05.

Simulation studies

Simulation setting

We performed simulation studies to examine the precision of the FPR estimated by the proposed method. In the simulation studies, the number of patients n is set to 200. The number of genes p is set to 1,000, including the p_1 (= 5, 30) outcome-predictive genes, ie, TP genes. The coefficient for gene j ($j = 1, \ldots, p$) β_j is set to 1.5 for the outcome-predictive genes ($j = 1, \ldots, p_1$) and 0 for the non-outcome-predictive genes ($j = p_1 + 1, \ldots, p$). The number of component C is set to 1 throughout. We may not be able to assume independence among genes, since the expression levels among the outcome-predictive genes are likely to be correlated because of gene co-regulation. It may be reasonable to assume that the

expression levels among the non-outcome-predictive genes as well as those between the outcome-predictive genes and the non-outcome-predictive genes are independent.[17] The gene expression levels for patient i, x_i, are generated from the multivariate normal distribution with mean vector 0 and covariance matrix Σ with variance 1, so that the correlation among the expression levels of the outcome-predictive genes is 0.0, 0.2, or 0.5, and is constant among the outcome-predictive genes. The survival time for patient i is generated on the basis of the exponential model as follows:

$$t_i = -\log(U)/\exp(x_i^T \beta) \quad (10)$$

where U is the uniform random variable between 0 and 1.[18] We set λ to 10–30 by 5 in the simulation studies in order to evaluate the precision of the estimated FPR for various values of λ, although the optimal value of λ is determined by cross validation in practice. The value of ζ is defined as the minimum value among $|\hat{\beta}_j|(\neq 0)(j = 1, \ldots, p)$ in the simulation studies. The average value for true FPR, the estimated numbers of both TP and FP genes, and the estimated FPR in 1,000 simulations are reported.

Simulation results

Table 1 shows that the average of the genes with $\hat{\beta}_j \neq 0$ in the lasso, true FPR, and the estimated TP, FP, and FPR for each design parameters in 1,000 simulations. According to Table 1, we found that the accuracy of the estimated FPR varied depending on the value of λ. Specifically, the accuracy of the estimated FPR was satisfactory for the values of $\lambda = 10$, 15, and 20, and it was slightly underestimated for the values of $\lambda = 25$ and 30. The number of genes with $\hat{\beta}_j$ was rela-

Table 1. Accuracy of the FPR estimated using the method proposed in the simulation studies.

ρ	p_1	λ	$\#\{j; \hat{\beta}_j \neq 0\}$	True FPR, %	$\widehat{\text{FPR}}$, %	$\widehat{\text{TP}}$	$\widehat{\text{FP}}$
0	5	10	126.2	96.0	96.0	5.0	121.1
		15	69.2	92.7	92.6	5.1	64.1
		20	31.0	83.5	82.9	5.2	25.8
		25	12.4	57.4	53.4	5.5	6.9
		30	6.5	19.8	14.6	5.4	1.1
	30	10	106.7	71.7	71.4	30.3	76.5
		15	72.1	57.7	56.8	30.6	41.5
		20	57.8	50.0	44.0	32.0	25.8
		25	42.5	44.4	32.5	28.5	14.1
		30	28.2	36.9	27.9	20.2	8.0
0.2	5	10	122.7	95.9	95.9	5.0	117.6
		15	65.4	92.3	92.2	5.1	60.3
		20	27.8	81.5	80.7	5.2	22.5
		25	10.3	48.6	43.8	5.5	4.8
		30	5.7	10.1	6.8	5.2	0.5
	30	10	64.1	52.8	52.0	30.5	33.6
		15	32.1	6.4	5.0	30.4	1.7
		20	30.0	0.1	0.1	30.0	0.0
		25	30.0	0.0	0.0	30.0	0.0
		30	30.0	0.0	0.0	30.0	0.0
0.5	5	10	119.3	95.8	95.8	5.0	114.2
		15	62.5	91.9	91.8	5.1	57.4
		20	25.4	79.7	78.8	5.2	20.2
		25	9.2	42.6	36.4	5.5	3.6
		30	5.4	6.5	3.3	5.2	0.2
	30	10	59.8	49.5	48.5	30.5	29.3
		15	31.1	3.4	2.1	30.4	0.7
		20	30.0	0.0	0.0	30.0	0.0
		25	30.0	0.0	0.0	30.0	0.0
		30	30.0	0.0	0.0	30.0	0.0

tively small for the larger value of λ; therefore, the degree of underestimation observed in the simulation studies may be acceptable. For instance, in case of $\rho = 0.0$, $p_1 = 5$, and $\lambda = 30$, the average number of true and estimated FP genes were 1.3 (= 6.5×0.198) and 1.0 (= 6.5×0.146), respectively, and the difference between them was negligibly small in practice. Furthermore, the values of ρ and p_1 did not greatly impact the accuracy of the FPR estimated.

Application to DLBCL data

We illustrated the exclusion of the FP genes from the genes selected by the lasso through the application of the proposed method to a real data set comprising the overall survival in 240 DLBCL patients with the expression of 7,399 genes.[1] The survival times were observed in 138 patients, and the censored times, in 102 patients. The median follow-up time was 3.9 years, and the median survival time was 2.8 years.

We divided the 240 patients into 2 groups; the training data comprised 160 patients, and the validation data, 80 patients, as described by Rosenwald et al.[1] We determined that the optimal value of λ was 27 by performing 5-fold cross validation, resulting in the selection of 12 genes as the outcome-predictive genes. Table 2 shows the GenBank accession number, description, and coefficient estimate for each of the 12 genes selected by the lasso.

Given the estimated coefficients $\hat{\beta}_j\,(j = 1, \ldots, 7399)$, we assume that the 2 mixture distributions with $C = 1$ and 2, and compared their fitness by using Akaike Information Criterion (AIC).[19] AIC is the most well known criterion for determining the number of components in the models. As a result, we selected the value of $C = 1$ for simplicity of interpretation, although the AICs for $C = 1$ and 2 were almost same. Thus, we assumed the mixture distribution with $C = 1$, and obtained the following estimated distribution (Fig. 1):

$$f\left(\hat{\beta}_j\right) = \frac{160}{7399}\left\{0.75 fL\left(\hat{\beta}_j; 0, 0.0053\right) + 0.25 f_N\left(\hat{\beta}_j; -0.10, 0.0064\right)\right\} + \frac{7239}{7399} f_L\left(\hat{\beta}_j; 0, 10^{-8}\right) \quad (11)$$

The mixed proportions of the Laplace and normal distributions in the *n/p* term were too small; therefore, we enlarged the part including these distributions in Figure 1. In addition, according to the estimated mixture distribution, the outcome-predictive genes that increase the risk of death, ie, genes with $\hat{\beta}_j > 0$, were not found.

Table 3 shows that the estimated numbers of FP and TP genes and the corresponding estimated FPR for various cut-off values. The estimated FPR was less than 5.0% for the cut-off value $\zeta > 0.03$, indicating that 3 genes might be TP genes, although the FPR might be underestimated according to the results of the simulation studies. In order to determine 9 genes that were most likely to be FP genes, we calculated the AICs of all possible models consisting of 3 genes selected among 12 genes, ie, 220 models in total. The model including

Table 2. The GenBank accession numbers, descriptions, and coefficient estimates of 12 genes selected by the lasso.

GenBank accession number	Description	$\hat{\beta}$
AA805575	Thyroxine-binding globulin precursor	–0.1039
X00452	Major histocompatibility complex, class II, DQ alpha 1	–0.1026
LC_29222	–	–0.0927
AF044323	COX15 homolog, cytochrome c oxidase assembly protein (yeast)	0.0167
L19872	Hydrocarbon receptor	–0.0078
M20430	Major histocompatibility complex, class II, DR beta 5	–0.0076
K01171	Major histocompatibility complex, class II, DR alpha	–0.0067
X59812 (R92015)	Cytochrome P450, subfamily XXVIIA polypeptide	–0.0028
M63438	Immunoglobulin kappa constant	0.0028
X82240 (AA729003)	T-cell leukemia/lymphoma 1A	–0.0017
X82240 (R97095)	T-cell leukemia/lymphoma 1A	–0.0010
X59812 (H98765)	Cytochrome P450, subfamily XXVIIA polypeptide	–0.0002

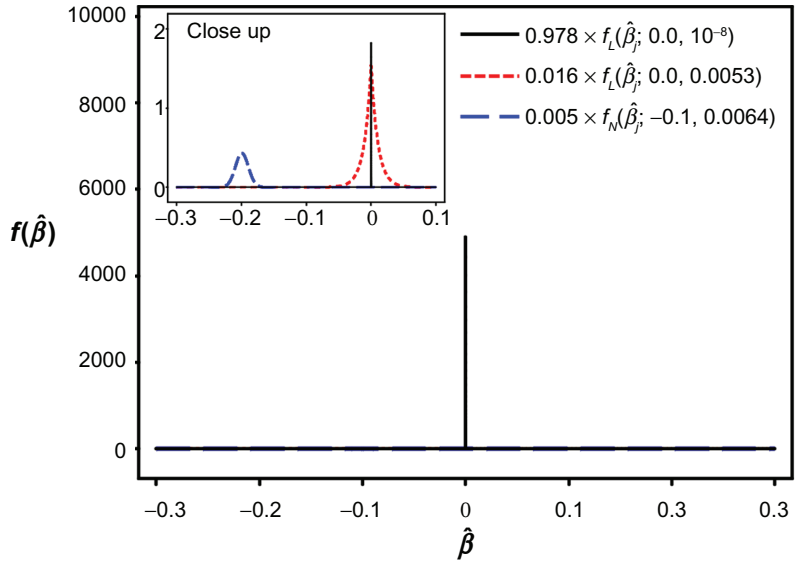

Figure 1. The estimated mixture distribution assuming the lasso estimates in the DLBCL data; f_L and f_N are the probability density functions of laplace and normal distributions, respectively. $\hat{\beta}$ is the estimate by the lasso and $f(\hat{\beta})$ is the probability density of $\hat{\beta}$.
Note: A magnified image of the distribution between the $\hat{\beta}$ values –0.3 and 0.1 is inserted.

3 genes with $\hat{\beta}$ values of –0.1039, –0.1026, and –0.0927 for AA805575, X00452, and LC_29222 showed the lowest AIC, and therefore, the remaining 9 genes were considered as FP genes.

Gene Set Enrichment Analysis

As an alternative method for the exclusion of the FP genes from the genes selected by the lasso, we used the Gene Set Enrichment Analysis (GSEA),[20] a computational method that assesses whether an *a priori* defined set of genes shows statistically significant relevance to survival time. The set of genes to be assessed by GSEA is generally defined based on the functional/biological relevance of gene expression profiles, such as genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same Gene Ontology (GO) category. In this study, for the application of the GSEA to the DLBCL data, we identified 1,454 sets of genes based on the GO categories. Of these, 53 gene sets included at least 1 of the 12 genes selected by the lasso method. It should be noted that 5 genes (eg, M20430, AA805575, M63438, LC_29222, and L19872) were not included in any of the gene sets. For this study, we implemented the modified GSEA for survival time proposed by Lee et al.[21] Table 4 shows 38 gene sets with false discovery rate (FDR) < 0.50 estimated by the modified GSEA. According to Table 4, the gene sets, including AF044323 and K01171, showed lower *P*-value and FDR, and therefore, we determined these genes as TP genes, and the remaining 10 genes were conveniently considered FP genes.

Table 3. The estimated numbers of TP and FP genes and the estimated FPR for the cut-off values from 0.0001 to 0.05.

Cut-off ζ	# $\{j; \lvert\hat{\beta}_j\rvert > \zeta\}$	$\widehat{FP}$	$\widehat{TP}$	$\widehat{FPR}$, %
0.0001	12	8.96	3.04	74.6
0.0005	11	8.05	2.95	73.2
0.001	10	7.13	2.87	71.3
0.005	7	3.76	3.24	53.7
0.01	4	1.24	2.76	30.9
0.02	3	0.19	2.81	6.3
0.03	3	0.03	2.97	1.0
0.04	3	0.00	3.00	0.0
0.05	3	0.00	3.00	0.0

Prediction accuracy

We demonstrated that the 9 genes identified did not impact the survival outcome by comparing the prediction accuracy between the models consisting of the aforementioned 3 and all 12 genes. Furthermore, we also compared the prediction accuracy between the models by which 3 TP genes were identified by the proposed method and 2 TP genes were identified

Table 4. Gene sets with FDR < 0.5 in the GSEA.

Gene set	*P*-value	FDR	The genes included in the gene set
Biosynthetic process	<0.001	<0.001	AF044323
Cellular biosynthetic process	<0.001	<0.001	AF044323
Mitochondrial part	0.002	0.035	AF044323
Mitochondrion	0.005	0.066	AF044323
Mitochondrial envelope	0.008	0.085	AF044323
Cytoplasmic part	0.014	0.093	AF044323, K01171
Lytic vacuole	0.014	0.093	K01171
Lysosome	0.014	0.093	K01171
Vacuole	0.022	0.103	K01171
Cellular component assembly	0.025	0.103	AF044323
Protein metabolic process	0.028	0.103	AF044323
Cellular macromolecule metabolic process	0.028	0.103	AF044323
Secondary metabolic	0.029	0.103	AF044323
Pigment biosynthetic process	0.029	0.103	AF044323
Pigment metabolic process	0.029	0.103	AF044323
Cellular protein metabolic process	0.034	0.109	AF044323
Mitochondrial membrane	0.035	0.109	AF044323
Cytoplasm	0.039	0.115	AF044323, K01171
Heme biosynthetic process	0.047	0.125	AF044323
Heme metabolic process	0.047	0.125	AF044323
Heterocycle metabolic process	0.067	0.169	AF044323
Macromolecular complex assembly	0.082	0.198	AF044323
Cofactor biosynthetic process	0.106	0.244	AF044323
Protein complex assembly	0.111	0.245	AF044323
Cofactor metabolic process	0.134	0.284	AF044323
Mitochondrial inner membrane	0.143	0.292	AF044323
Receptor activity	0.184	0.349	X00452
Multicellular organismal development	0.191	0.349	X82240
Transmembrane receptor activity	0.191	0.349	X00452
Organelle inner membrane	0.200	0.349	AF044323
Cellular protein complex assembly	0.209	0.349	AF044323
Envelope	0.217	0.349	AF044323
Organelle envelope	0.217	0.349	AF044323
Organelle part	0.324	0.467	AF044323
Intracellular organelle part	0.324	0.467	AF044323
Inorganic cation transmembrane transporter activity	0.324	0.467	AF044323
Mitochondrial membrane part	0.326	0.467	AF044323
Cytochrome c oxidase activity	0.356	0.497	AF044323

Table 5. Three criteria for model evaluation.

Criteria	Model with 3 genes identified by the proposed method	Model with 12 genes	Model with 2 genes identified by the GSEA
P-value of the log-rank test	0.002	0.007	0.246
P-value of the prognostic index	0.002	0.002	0.120
Deviance	–8.942	–9.072	–1.967

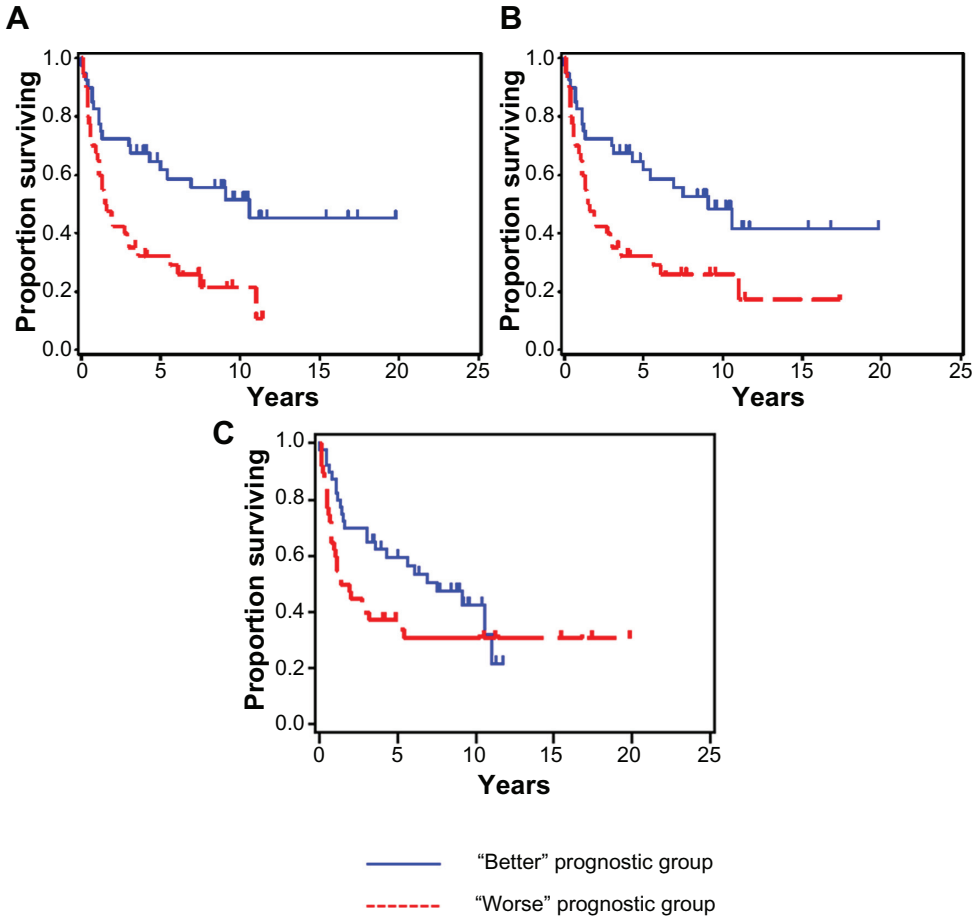

Figure 2. Kaplan-Meier curves of overall survival for the 2 groups; (**A**) in the models that identified 3 genes by the proposed method, (**B**) in the models that identified 12 genes by the lasso method, (**C**) in the models that identified 2 genes by the GSEA.

by the GSEA. For the validation data including 80 patients, the following 3 criteria were calculated: *P*-value for the log-rank test, *P*-value for the prognostic index, and deviance. The 80 patients were categorized into 2 groups by the boundary of the median of prognostic index $\hat{\eta}_i = \mathrm{x}_i^{\mathrm{T}}\beta$; the "better" and "worse" prognostic groups. The Kaplan-Meier curves between the 2 groups were compared by the log-rank test. Next, we calculated the *P*-value for the parameter α multiplied by the prognostic index $\hat{\eta}_i$ in the Cox proportional hazard model $h(t_i \mid \mathrm{x}) = h_0(t)\exp(\alpha\hat{\eta}_i)$. Finally, the deviance was calculated by $-2\{l^{(validation)}(\hat{\beta}_{training}) - l^{(validation)}(0)\}$ where $l^{(validation)}(\hat{\beta}_{training})$ and $l^{(validation)}(0)$ are the Cox log partial-likelihood function for the estimated coefficients by using training data and zero vector 0, respectively. For each criterion, the smaller value suggests better prediction accuracy. The values of the 3 indices for the 3 models—the proposed method that identified 3 TP genes, the lasso method that identified 12 genes, and the GSEA that identified 2 TP genes—are shown in Table 5. As shown in Table 5, the values of the 3 indices between the models that identified 3 and 12 TP genes are almost the same. Furthermore, the prediction accuracy of the proposed method was found to be better than that of the GSEA. In addition, the Kaplan-Meier curves of the overall survival for the proposed and lasso methods were also similar (Fig. 2). Figure 2 shows that the difference in the overall survival between the 2 groups was significant for the proposed method, but not for the modified GSEA. Thus, by using the proposed method, we are able to exclude the genes that are not likely to impact the survival outcome.

Discussions and Conclusions

In this study, we developed a method to estimate FPR by assuming the mixture distribution comprising the Laplace and normal distributions on the lasso estimates. In practice, we identified the

outcome-predictive genes by performing the lasso, and subsequently, removing the FP genes using the proposed method.

Although the penalized regression analyses including the lasso are attractive in the high-dimensional microarray data, it is difficult to identify the outcome-predictive genes without FP genes by using these methods. Utilizing the proposed method, we can validate the results of the lasso, and identify the outcome-predictive genes more precisely. The assumed mixture distribution was formulated considering the 2 features of the lasso, although it may be a "somewhat complex" distribution. The validity of this assumption was demonstrated through the simulation studies. Specifically, the accuracy of the FPR estimated by the proposed method was satisfactory in many cases. The accuracy was slightly decreased for the larger value of tuning parameter λ, but the underestimation of FPR may be acceptable in practice, as discussed in the Simulation section.

In the section on Application to the DLBCL Data, the utility of the proposed method was illustrated. We were able to eliminate the FP genes from the genes selected by the lasso with $\lambda = 27$, and improved the accuracy of prediction of the model. We further identified the TP genes and examined the prediction accuracy of overall survival based on them, using the proposed method and GSEA. Both methods identified no TP genes in common. The prediction accuracy using the 3 genes identified by the proposed method outperformed that using the 2 genes identified by the GSEA. The GSEA introduced by the Subramanian et al[20] evaluates microarray data at the level of gene sets. The gene sets are defined based on prior biological knowledge, eg, published information about biochemical pathways or coexpression in previous experiments. In contrast, the proposed method evaluates microarray data at the level of genes and does not use prior biological knowledge when identifying the outcome-predictive genes.

Some variants of the lasso and penalized regression methods are used, eg, smoothly clipped absolute deviation penalty (SCAD),[22] adaptive lasso,[23,24] elastic net,[16] and ridge regression,[25] but of these, we chose the lasso in this study, because of our concerns regarding the high possibility of missing the true positives for the SCAD and adaptive Lasso, the difficulty in choosing 2 penalties for the elastic net, and the absence of any property to select genes for ridge regression.[10]

The determination of the value of the tuning parameter is required when performing the lasso. The value of λ is frequently determined on the basis of the cross validation that evaluates the adequacy of the model, as explained in the Methods section. By utilizing the proposed method, we could also determine the value of λ by considering not only the prediction accuracy but also the FPR.

In conclusion, the lasso allows us to efficiently select the outcome-predictive genes in the high-dimensional microarray data, but the difficulty lies in the inclusion of the FP genes among the selected genes. The use of the proposed method allows us to eliminate these genes and improve the prediction accuracy of the Cox model.

Author Contributions

Conceived and designed the experiments: SK, AH. Analysed the data: SK, AH. Wrote the first draft of the manuscript: SK, AH. Contributed to the writing of the manuscript: SK, AH, CH. Agree with manuscript results and conclusions: SK, AH, CH. Jointly developed the structure and arguments for the paper: SK, AH, CH. Made critical revisions and approved final version: SK, AH, CH. All authors reviewed and approved of the final manuscript.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Rosenwald M, Wright G, Chan CW, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*. 2002;346:1937–47.
2. Van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347: 1999–2009.
3. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6.
4. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005; 365:671–9.
5. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*. 2007;99:147–57.
6. Ein-Dor L, Kela I, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*. 2005;21:178–8.
7. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc, Series B Stat Methodol*. 1996;58:267–88.
8. Tibshirani R. The Lasso method for variable selection in the Cox model. *Stat Med*. 1997;16:385–95.
9. Bøvelstad HM, Nygård S, Størvold HL, et al. Predicting survival from microarray data—a comparative study. *Bioinformatics*. 2007;23:2080–7.
10. Benner A, Zuchnick M, Hielscher T, Ittrich C, Mansmann U. High-dimensional Cox model: The choice of penalty as part of the model building process. *Biom J*. 2010;52:50–69.
11. van Wieringen WN, Kun D, Hampel R, Boulesteix AL. Survival prediction using gene expression data: a review and comparison. *Comput Stat Data Anal*. 2009;53:1590–603.
12. Verweij PJ, Houwelingen HC. Cross-validation in survival analysis. *Stat Med*. 1993;12:2305–14.
13. Cox DR. Regression models and life-table (with disscusion). *J R Stat Soc, Series B Stat Methodol*. 1972;74:187–220.
14. Goeman J. L1 penalized estimation in the Cox proportional hazards model. *Biom J*. 2010;52:70–84.
15. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32:407–51.
16. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc, Series B Stat Methodol*. 2005;67:301–20.
17. Tsai CA, Wang SJ, Chen DT, Chen JJ. Sample size for gene expression microarray experiments. *Bioinformatics*. 2005;21:1502–8.
18. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2006;24:1713–23.
19. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19:716–23.
20. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545–50.
21. Lee S, Kim J, Lee S. A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC Bioinformatics*. 2011;12:377.50.
22. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96:1348–60.
23. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006; 101:1418–29.
24. Zhang H, Lu W. Adaptive lasso for Cox's proportional hazard model. *Biometrika*. 2007;94:691–703.
25. Hoerl AE, Kennard RW. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*. 1970;35:109–47.