Journal of
Health & Medical Informatics

Kitamura et al., J Health Med Informat 2015, 6:1
DOI: 10.4172/2157-7420.1000180

Research Article Open Access

Statistical Properties and Power Analysis of Cox's Proportional Hazards Model Regularized by Various Penalties for DNA Microarray Gene Expression Survival Data

Nobutaka Kitamura[1*], Kouhei Akazawa[1] and Kosuke Yoshihara[2]

[1]Department of Medical Informatics, Niigata University Medical and Dental Hospital, Niigata 951-8520, Japan
[2]Department of Obstetrics and Gynecology, Niigata University Graduate School of Medical and Dental Sciences, Niigata 951-8520, Japan

Abstract

Background: Compared with the number of candidate genes used for DNA microarray experiments, the number of available samples is extremely limited. As a result, overfitting of the data may occur during regression analyses. To solve this problem, various penalized regression models have been suggested. In general, the validity of a regression model should be verified using a validation data set, as opposed to the training data used to construct the model. However, at present there are no programs available to calculate statistical properties, including the precision, validity, and the statistical power of the Cox's proportional hazards model regularized by various penalties; therefore, the properties of these models are not sufficiently clear.

Methods: In this study, we created programs using the R language to calculate statistical properties of the Cox's proportional hazards model, including the statistical power based on the prognostic index, and conducted simulation experiments under various conditions of DNA microarray expression data with survival time.

Results: The results showed that the power of a validation set for penalized methods is greater than for stepwise methods in many cases, particularly when $n < p$. This tendency is most remarkable for the penalized methods including both the L1-norm and the L2-norm. Furthermore, we tested our programs using actual microarray gene expression data with survival time data to confirm their validity.

Conclusions: Our simulation programs for the Cox's proportional hazards model regularized by various penalties are very useful for planning DNA microarray studies or for evaluating the results of such studies.

Keywords: Penalized Cox's proportional hazards model; LASSO; Elastic net method; Ridge method; Statistical power; Validity

Introduction

DNA microarray experiments are commonly used to identify disease susceptibility genes efficiently. The experimental system for microarrays has progressed remarkably, and it has now become possible to measure the fluorescence intensities of hundreds of thousands of genes simultaneously. Genome-wide expression profiles that assess risk factors for a disease offer the possibility of more precisely defining clinical prognosis; however, compared with the number of candidate genes that can be used for DNA microarray experiments, the numbers of such expression profiles that are available are extremely limited. This situation may lead to a rise in the number false positive statistical tests across genome-wide association studies.

In regression analyses, the precision of estimated parameters may decrease and overfitting of data may occur, yielding a regression model with poor predictive power and accuracy that cannot be used for other data sets [1]. As a result, researchers have tried and failed to identify robust and highly accurate prognostic biomarkers. To help address this problem, various multiple comparison methods or sequential step-by-step procedures have been applied to association studies, and several penalized regression models [2-5], such as the ridge model, the lasso method, and the elastic net method, have been advocated for regression analyses. Tibshirani et al. [2,4] developed an algorithm and R programs for the Cox's proportional hazards model regularized by various penalties. In general, the validity of a regression model should be verified using a validation data set rather than the training data that was used to construct the regression model. However, at present no programs are available for calculating statistical properties, such as the statistical power of the Cox's proportional hazards model regularized by various penalties, so the statistical properties of these regression models are not sufficiently clear.

In this study, we created programs in the R language to calculate statistical properties, including the statistical power based on the prognostic index of the Cox's proportional hazards model regularized by various penalties and on the prognostic index of the stepwise method. Using these programs, we conducted simulation studies for DNA microarray experiments under various conditions and compared various penalized regression models with a stepwise Cox regression analysis. Furthermore, the optimal solving method of the two-dimensional parameters (the tuning parameter λ and the mixing parameter α) of the penalized proportional hazards model is still controversial. Here, we determined these parameters using the prognostic indices mentioned above, and verified the reproducibility by actual gene expression survival data using these methods.

***Corresponding author:** Nobutaka Kitamura, Department of Medical Informatics, Niigata University, Medical and Dental Hospital, 1-754 Chuo-ku, Asahimachi-dori, Niigata 951-8520, Japan, Tel: +81252272471; E-mail: nktmr@m12.alpha-net.ne.jp

Received January 07, 2015; **Accepted** January 27, 2015; **Published** February 02, 2015

Citation: Kitamura N, Akazawa K, Yoshihara K (2015) Statistical Properties and Power Analysis of Cox's Proportional Hazards Model Regularized by Various Penalties for DNA Microarray Gene Expression Survival Data. J Health Med Informat 6: 180. doi:10.4172/2157-7420.1000180

Methods

Cox's proportional hazards model regularized by various penalties

Let t be a survival time, $z=(z_1,\cdots,z_p)'$ a vector of predictors, $r(z_1,\cdots,z_p)=r(z)$ a function of z, $h(t\mid z)$ a hazard at time t given predictor z, and $h_0(t)$ an arbitrary baseline positive valued hazard function at time t. Then, the Cox's proportional hazards model can be represented as:

$$h(t\mid z_1,\cdots,z_p)=h_0(t)r(z_1,\cdots,z_p)=h_0(t)r(z)\quad (t>0).$$

If we assume $r(z)$ to be

$$r(z_1,\cdots,z_p)=\exp(\beta_1 z_1+\beta_2 z_2+,\cdots,+\beta_p z_p)=\exp(\boldsymbol{\beta}'z)$$

where $\boldsymbol{\beta}'=(\beta_1,\cdots,\beta_p)$ indicates a parameter vector of the coefficients of predictors.

Assume that for each case from 1, …, n, we have predictors $z_1,\ldots,z_n$ and survival times $t_1,\ldots,t_n$, including censoring times, and suppose that k out of n cases die. Let $t_{(1)}<t_{(2)}<\cdots<t_{(k)}$ be an increasing list of failure times and $z_{(1)},\ldots,z_{(k)}$ be a sequence of corresponding predictors where $(1),\ldots,(k)$ index the number of observations. Let $R(t_{(i)})$ be a risk set just before $t_{(i)}$. Inference is then made via the partial likelihood as:

$$L(\beta)=\prod_{i=1}^{k}\left(\frac{\exp(\beta' z_{(i)})}{\Sigma_{l\in R(t_{(i)})}\exp(\beta' z_l)}\right),$$

and the log partial likelihood is

$$\frac{2}{n}l(\beta)=\frac{2}{n}\left[\beta' z_{(i)}-\log\left(\Sigma_{l\in R(t_{(i)})}\exp(\beta' z_l)\right)\right]$$

Where, for convenience, $2/n$ is used to modify the equations.

The constraint using the L1-norm and the L2-norm, where α ($0\le\alpha\le 1$) and (1-α) are mixing parameters for the L1-norm and the L2-norm, respectively, is as follows:

$$\alpha\sum_{i=1}^{k}|\beta_i|+(1-\alpha)\sum_{i=1}^{k}\beta_i^2\le c,$$

and the penalty term $\lambda P_\alpha(\beta)$ is determined using the constraint and the tuning parameter λ:

$$\lambda P_\alpha(\beta)=\lambda\left(\alpha\sum_{i=1}^{p}|\beta_i|+(1-\alpha)\sum_{i=1}^{p}\beta_i^2\right)$$

Hence, considering the Lagrangian formulation, the estimate of coefficient β at a λ is calculated as:

$$\hat{\beta}=\arg\max_\beta\left[\frac{2}{n}\left(\sum_{i=1}^{k}\beta' z_{(i)}-\log\left(\Sigma_{l\in R(t_{(i)})}\exp(\beta' z_l)\right)\right)-\lambda P_\alpha(\beta)\right]$$

Where $\alpha=0$ is the ridge penalty, $0<\alpha<1$ is the elastic net penalty, and $\alpha=1$ is the lasso penalty.

Because it is impossible to solve the above formula analytically, numerical calculations are performed using algorithms from non-convex optimization theory and the optimal value of λ is determined using a cross-validation method [4].

Algorithm for simulation study of various penalized Cox's proportional hazards models

1. Set conditions on the sample sizes (for the training set or validation set), the number of true disease susceptibility genes, the number of candidate genes, the coefficients for each gene, and other parameters for the exponential distribution. The microarray fluorescence intensities of each gene are chosen from an arbitrary distribution. In this study, we used the multivariate standard normal distribution with the correlation coefficient between each gene.

2. Generate random survival times (T) from the Cox's proportional hazards model by the reverse function method for a training set. The reverse function method is performed by setting $T=-r\cdot\log(U)\Big/\exp\left(\sum_{k=1}^{p}\beta_k z_k\right)$, where r is a mean of a probability density function and U is a uniform random number. Then, generate uniformly distributed random censored times and set a follow-up period. Let the minimums of those three periods (T, the uniformly distributed random censored times, and r) be an observation period, and let "1" be an endpoint time and "0" be a censored time.

3. Use the training set to calculate estimates of coefficients and to select genes by the stepwise Cox regression model and by various penalized regression models including the ridge method, the lasso method, and the elastic net method. In the stepwise regression analysis, the candidates that achieve statistical significance in the univariate Cox regression analyses are used subsequently in a stepwise multivariate regression analysis.

4. Generate another set of random survival times from Cox's proportional hazards model for a validation set. To calculate the statistical power of the Cox model, use the information about the selected genes and estimated values of coefficients from the training set to calculate the prognostic index. The prognostic index is defined as the sum of the product of coefficients and fluorescence intensity values of selected genes for each case in the validation set.

5. Divide the validation set into a high-risk group and a low-risk group using the median of the prognostic indices and conduct a log-rank test between the two groups.

6. Repeat the above procedure 2000 times. Calculate the mean coefficient of the selected genes, the mean number of selected genes, the true positive rate (TPR, the number of true genes among the selected genes divided by the number of true genes), the true negative rate (TNR, the number of no-true genes among the selected genes divided by the number of no-true genes), the positive predictive value (PPV, the number of true genes among the selected genes divided by the number of selected genes), the negative predictive value (NPV, the number of no-true genes among the selected genes divided by the number of unselected genes), and the statistical power (the proportion of p values less than 0.05) for the training and validation sets.

7. For the simulation experiments, we first set the number of true genes to 10, the true values of the coefficients to 0.2 and 0.4, the numbers of candidate genes to 100 and 1000, the values of the correlation coefficients to 0, and the sample sizes to 100, 150, and 200, and the survival time data as uncensored data sets and censored data sets. We set the censoring pattern as type I (meaning the censored time of the respective case is predetermined at the observation start time of each case) and set r and the follow-up period to 200 (the censor rate is set to 0.4). Next, we set the values of correlation coefficients to 0.1, 0.3, 0.5, and 0.7, the numbers of candidate genes (= p) to 100 and 1000, and the sample size (= n) to 100 and 200.

The above-mentioned programs were created using the R language version 3.1.1 and are available from the authors on request.

Results and Discussion

Simulated data sets

Results when the survival data were complete and the correlation coefficient was zero: The statistical powers of all the methods increased as both the sample sizes and the true values of coefficients increased, and the statistical powers decreased as the numbers of candidate genes increased. The statistical powers of the training sets for the stepwise methods and the ridge methods were the largest at 100%, followed by the elastic net methods and the lasso methods. However, the statistical powers of the validation sets were smaller compared with the statistical powers of the training sets for all the methods tested; in many cases, particularly when n < p, the powers of the validation sets for the elastic net method and the lasso method were the largest, followed by the ridge method and the stepwise methods (Figures 1 and 2).

Results when the survival data included censored times and the correlation coefficient was 0.1–0.7: When the data included censored cases, the statistical powers for each method decreased compared with uncensored data, but the shapes of the graphs were similar to the shapes of the graphs obtained using uncensored data (Figures 3 and 4).

When the correlation coefficients between each gene increased, the statistical powers for the penalized methods increased, and the powers for the stepwise methods decreased (Figure 5).

The results for the number of selected genes, coefficient of selected genes, standard deviation (SD) of coefficients for selected genes, TPR, and PPV are available as online supplementary material (Figures S1–S12).

Actual data set

Serous ovarian cancer tumors are heterogeneous; therefore, it is necessary to classify ovarian cancers appropriately based on their biological characteristics so that the medical treatment can match the character of each serous ovarian cancer. DNA microarray analysis can be a useful tool for diagnosis and prognostic predictions for many diseases, but because there is a scarcity of available ovarian cancer samples compared with the number of candidate genes, it is difficult to identify gene expression signatures for serous ovarian cancers with high accuracy and reproducibility.

Microarray data sets for ovarian cancer

Three actual microarray data sets derived from studies of serous ovarian cancer patients were obtained: the cancer genome atlas (TCGA) data set [6], Tothill's data set [7] and Bonome's data set [8]. These data sets are from microarray experiments that were set up to

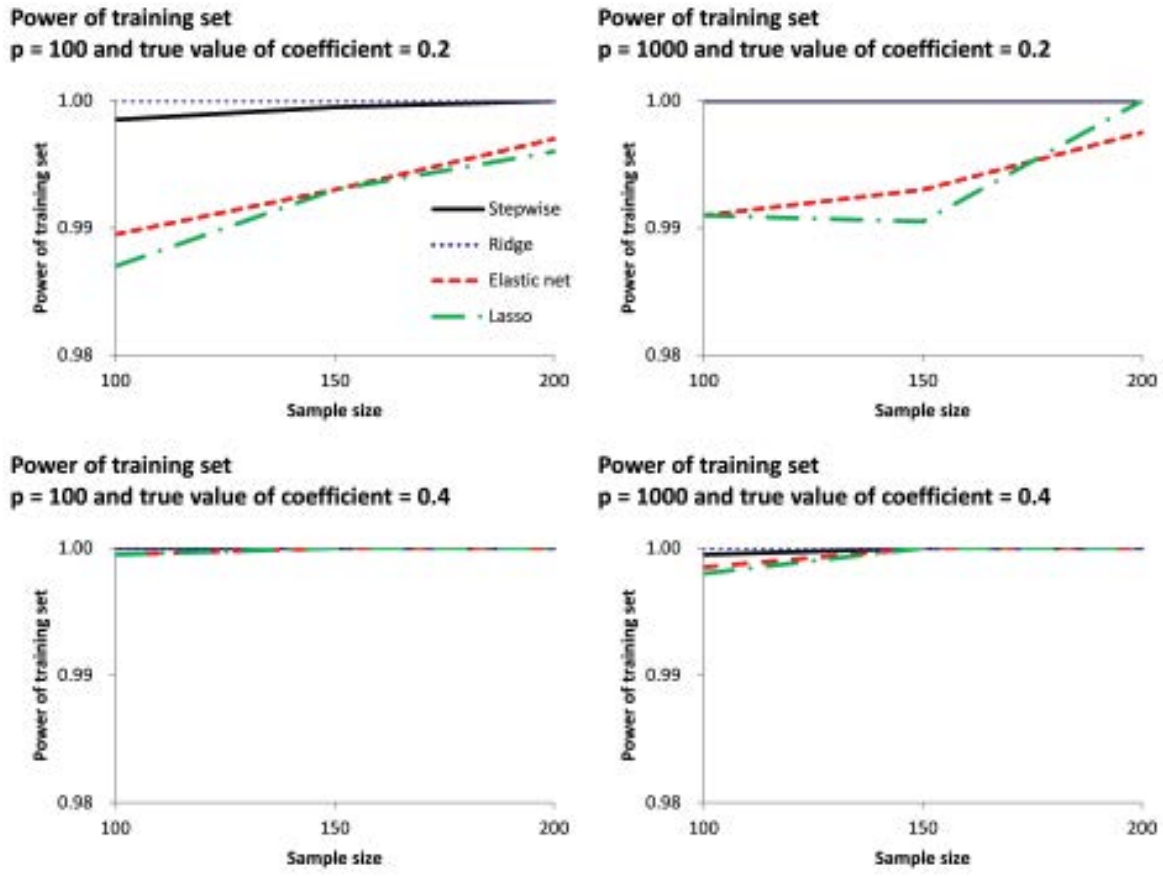

Figure 1: Statistical power of the training sets for the stepwise method and each penalized method using uncensored data sets. Numbers of candidate genes p = 100 and 1000, and true values of the coefficients = 0.2 and 0.4.

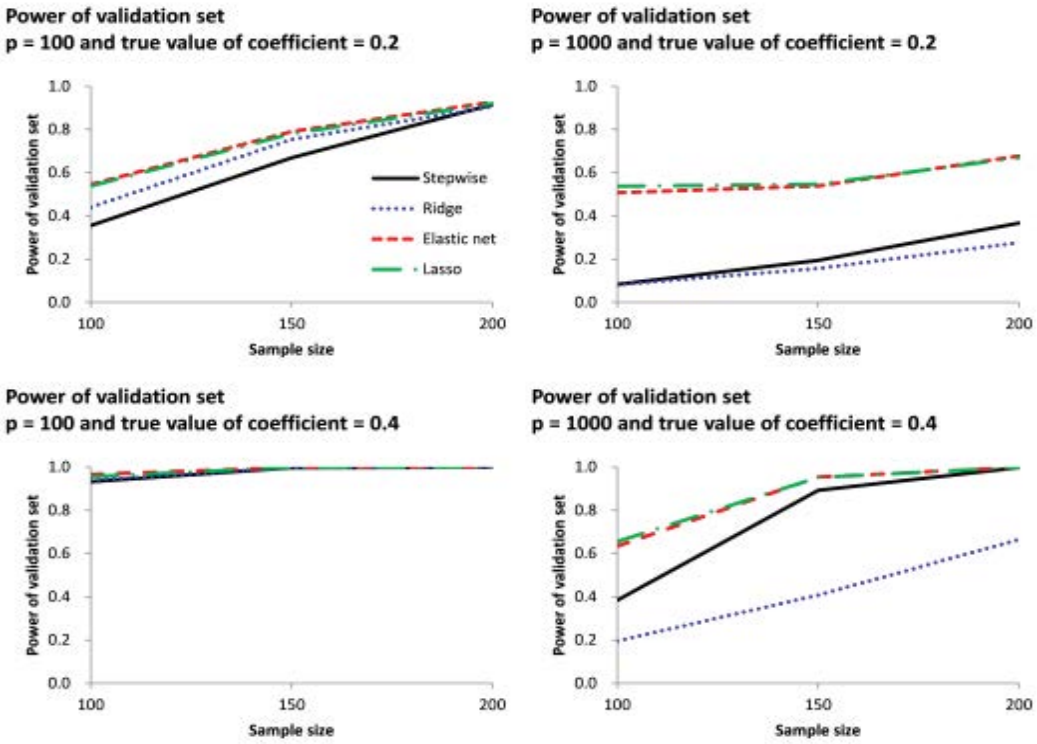

Figure 2: Statistical power of the validation set for the stepwise method and each penalized method using uncensored data sets. Numbers of candidate genes p = 100 and 1000, and true values of the coefficients = 0.2 and 0.4.

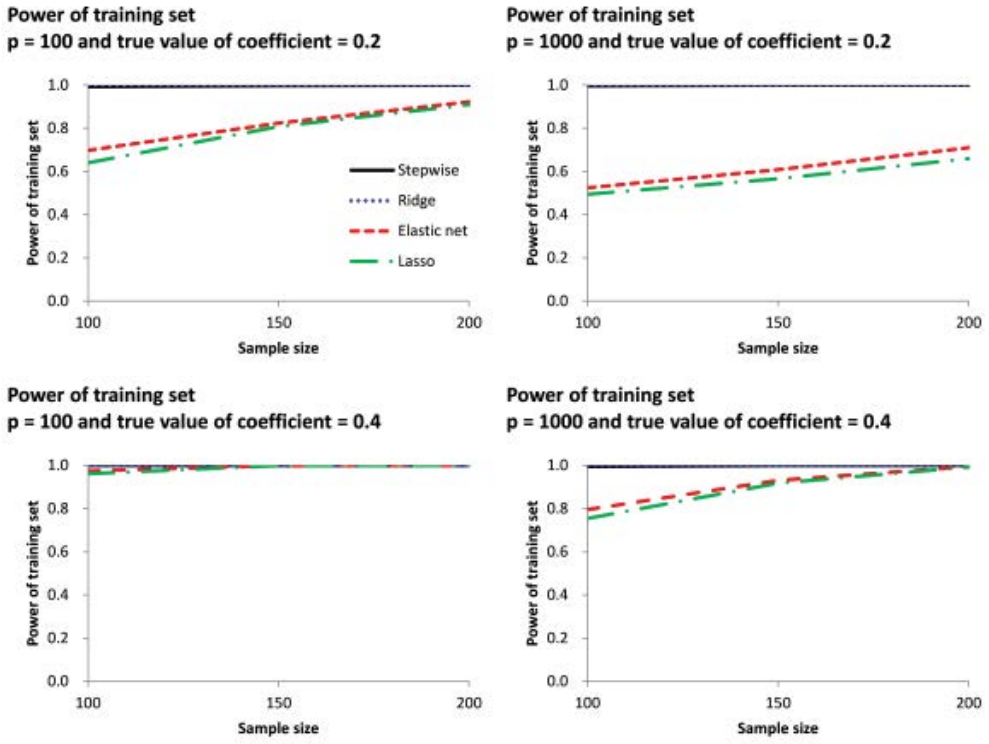

Figure 3: Statistical power of the training sets for the stepwise method and each penalized method using censored data sets. Numbers of candidate genes p = 100 and 1000, and true value of the coefficients = 0.2 and 0.4.

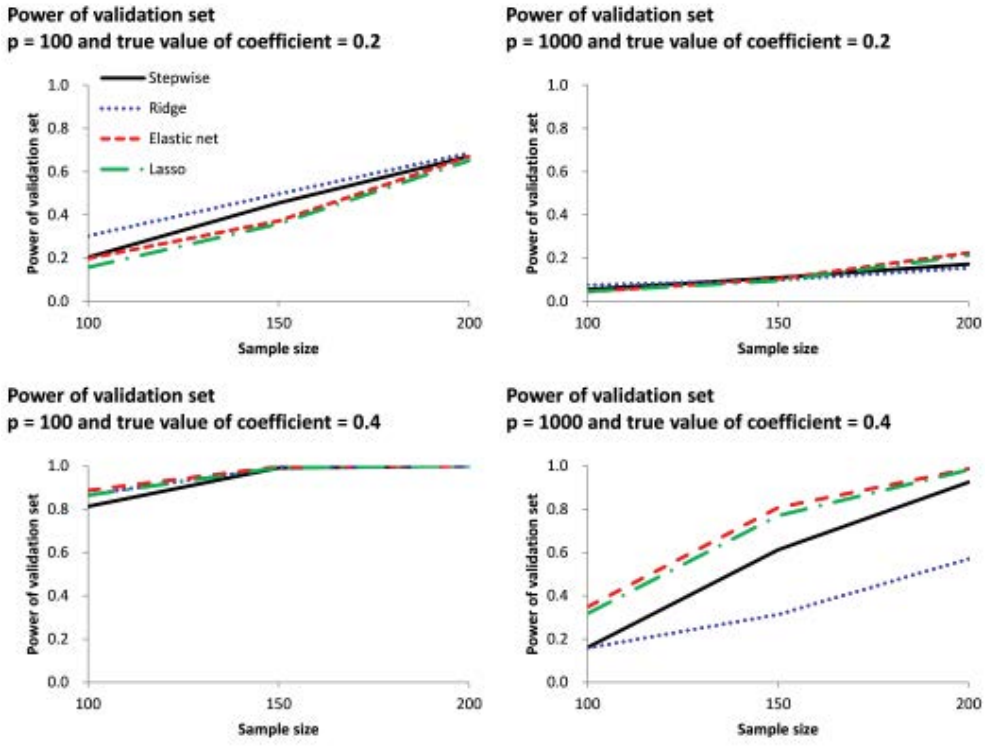

Figure 4: Statistical power of the validation set for the stepwise method and each penalized method using censored data sets. Numbers of candidate genes p = 100 and 1000, and true value of the coefficients = 0.2 and 0.4.

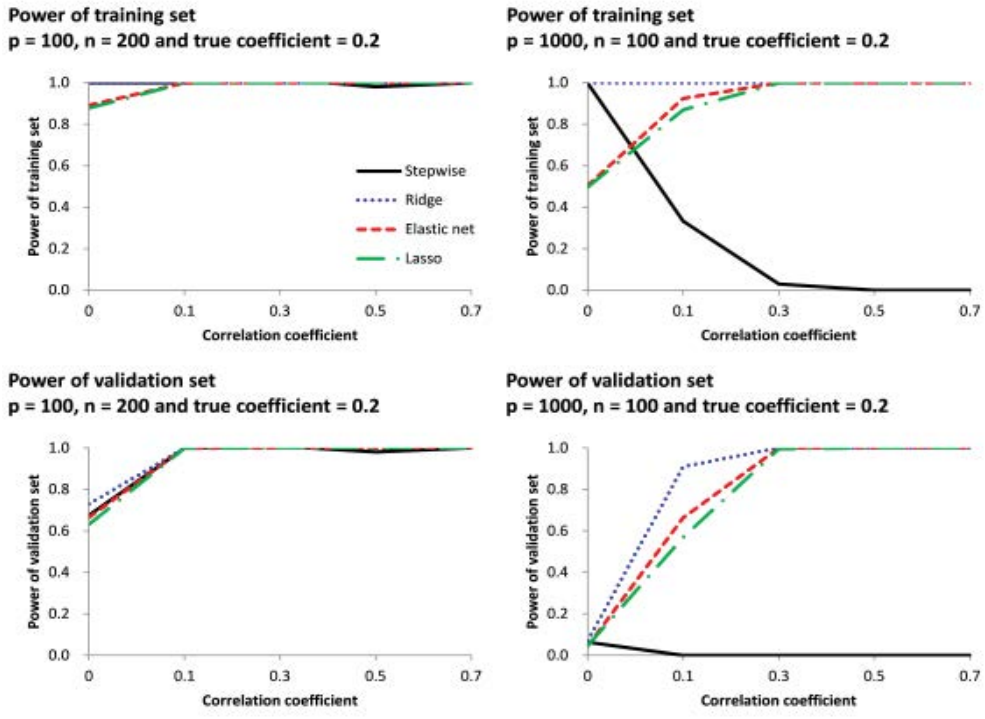

Figure 5: Statistical power of the training and validation sets for the stepwise method and each penalized method using censored data sets with correlation coefficients. When the numbers of candidate genes p = 100, n = 200 (n > p; first column), and when p = 1000, n = 100 (n < p; second column), and true value of the coefficient = 0.2.

investigate the survival of ovarian cancer patients. The data sets contain microarray expression data on a total of 1686 genes. The TCGA data set contains survival time data for 319 patients (follow-up period 1–154 months; type I censor rate 0.41), Tothill's data set contains survival time data for 131 patients (follow-up period 6–79 months; type I censor rate 0.54), and Bonome's data set contains survival time data for 185 patients (follow-up period 1–164 months; type I censor rate 0.30).

Reproducibility of the survival gene expression signatures

We verified the reproducibility of the survival gene expression signatures by comparing the statistical indices from the three actual data sets using the penalized Cox's proportional hazards model.

To choose the optimal mixing parameter α from the range 0–1, we calculated TPR, TNR, PPV, NPV, and statistical powers using our programs. We set the number of true genes to 100, the coefficient values to 0.7, the correlation coefficients between each gene to 0.01, the numbers of candidate genes to 1686, and the sample sizes to 100, 150, 200 and 250. The survival data were set to include censored time and the follow-up period was set to 200. (These values have been reported previously [6-8].)

The penalty term for the Cox's proportional hazards model has two parameters, the tuning parameter λ and the mixing parameter α described in methods. To determine the λ and α parameters, the same fold was set in separate calls by setting the fold ID number to fix in the program, and a two-dimensional cross-validation with different values of α was conducted. A graph of α versus the statistical indices was drawn with α set between 0.0 and 1.0 (at intervals of 0.05) on the horizontal axis and the statistical indices on the vertical axis). The optimal mixing parameter α was selected as the point where the respective statistical indices crossed each other on the graph.

The simulation showed that when the sample size was 200, the optimal mixing parameter α should be under 0.05 to detect coefficients of 0.7 or larger with at least 70% statistical power. When the sample size was 250, α should be under 0.05 to detect coefficients of 0.7 or larger with at least 80% power at the 5% level of significance (Figure 6, red arrow).

Therefore, we conducted an elastic net analysis on the TCGA data set (training set, n = 319) using α set at 0.05 and detected 25 genes as prognostic factors for high-grade serous ovarian cancers. Next the predictive set of 25 candidate genes was tested with the Tothill's data set (validation set, n = 185). Both the Kaplan–Meier method and the log-rank test showed that this set of 25 genes was significantly associated with overall survival time in the Tothill's data set (p value = 0.009, log-rank test). Furthermore, when the set of 25 genes was tested using Bonome's data set (validation set, n = 131 similar results were obtained (p value = 0.049, log-rank test) (Figure 7). The survival gene expression signatures from the training set detected by the penalized Cox's proportional hazards model showed good reproducibility with these independent validation sets.

Regression model analysis is used often for prognostic predictions with DNA microarray data. In such cases, there may be a large input dimensionality and a paucity of patient samples. This means that overfitting of data may be a persistent problem, often referred to as the "p >> n" problem.

Sparse estimation methods have attracted attention in recent years as a way for solving the p >> n problem. Solutions of the resulting simultaneous equation are found by regularization under the constraint that the number of solutions is sparse. The recommended constraints for the least squares method are the L1-norm (sum of absolute values of each coefficient) or the L2-norm (sum of squares of each coefficient), or a linear combination of the L1-norm and the L2-norm weighted by the mixing parameters α ($0 \le \alpha \le 1$) and 1-α. The constraint is weighted by the tuning parameter λ, called the penalty term, combined with the traditional formula of the sum of square errors. In the penalized Cox's proportional hazards model, the formula, which consists of the log partial likelihood function, and the penalty term are maximized simultaneously to achieve a shrunken estimation of the coefficients and to select the variables.

Solving this formula for a penalized model analytically is impossible, so numerical calculations using methods from non-convex optimization theory need to be used. Therefore, to clarify the statistical properties of the various penalized models, simulations are indispensable. Currently, no programs to calculate properties, such as the statistical power of Cox's proportional hazards models regularized by various penalties, are available. Therefore, the statistical properties of such models, including the predictive validity of the Cox's proportional hazards model or the penalized model, are not clear at present.

Furthermore, when a stepwise procedure using a very large number of candidate factors is performed, the calculation load for the stepwise procedure increases remarkably, making the calculation impossible. Therefore, the number of factors needs to be narrowed down with a series of univariate analyses so that a multivariate analysis can be conducted using a smaller list of factors. When a stepwise analysis was performed in our programs, the candidates that achieved statistical significance in the univariate Cox's regression analyses were used subsequently in a stepwise multivariate regression analysis. A comparison of the statistical properties of such a stepwise method after univariate analyses with various penalized methods has not been performed until now.

In this study, we created programs using the R language to calculate various indices for the statistical validation of Cox models, including the number of selected genes, the mean coefficient of selected genes, mean SD of selected genes, TPR, TNR, PPV, NPV, and statistical power. We compared various penalized regression models with the conventional stepwise multivariate Cox regression analysis.

When the sample size was larger than the number of candidate genes, the number of selected genes for the stepwise method approached the true value asymptotically as the sample size and the true value of the coefficients increased, but when the number of candidate genes was larger than the sample size, the number of selected genes was much larger than the true value, indicating that the stepwise method was unable to narrow down many candidates to a smaller number (see Figures S1 and S6). However, the number of selected genes for two of the penalized methods (lasso and elastic net) was much smaller than for the stepwise method (see Figures S1 and S6).

Tibshirani et al. [2] reported the results of a simulation study using 50 datasets, which showed that the coefficients and mean squared errors for the lasso method using L1-norm in Cox's model were smaller than those for the stepwise model. We observed a similar result in our simulations (see Figures S2, S3, S7, and S8). Furthermore, the simulation experiments using our programs indicated that the shrinkage effect of the coefficient and the SDs for L2-norm were higher than L1-norm. Thus, the L1-norm shrank the number of selected genes while L2-norm did not have the shrinkage effect, meaning that L1-norm could shrink the coefficient towards zero. A linear combination of L1-norm

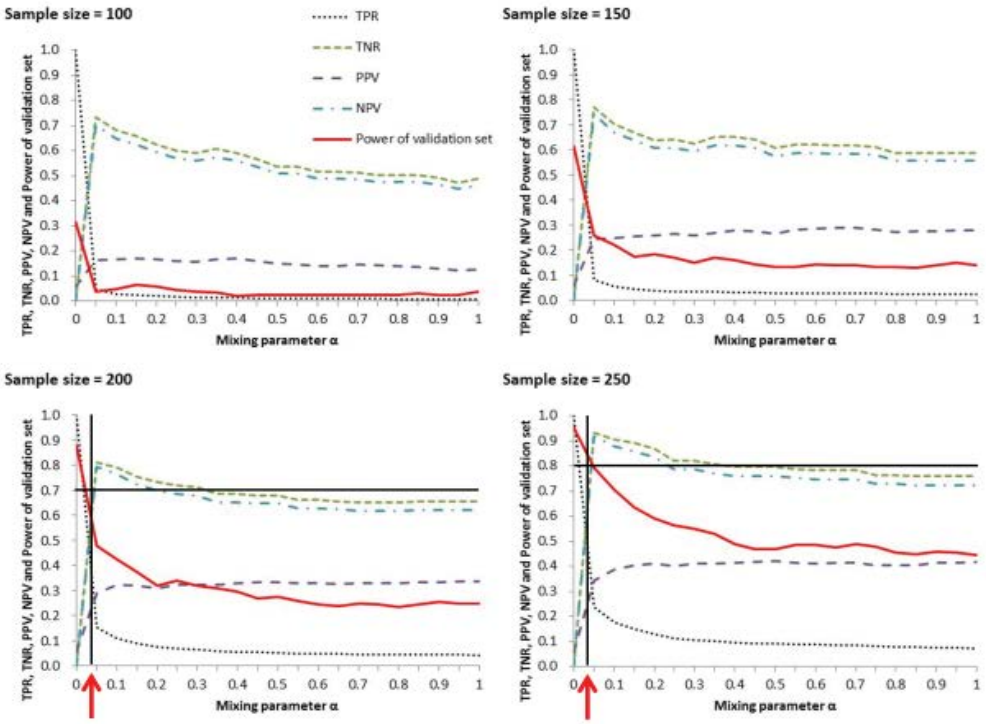

Figure 6: Statistical power and TPR, TNR, PPV, and NPV of the validation set for the penalized Cox's proportion hazard model using actual microarray data sets for ovarian cancer with mixing parameter α. The red arrow indicates α values under 0.05 can detect coefficients of 0.7 or larger with at least 80% power at the 5% significance level.

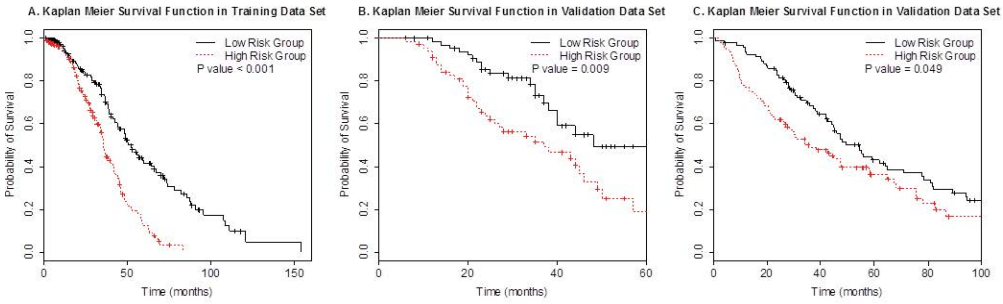

Figure 7: Verification of the validity of the penalized Cox's proportional hazards model using actual microarray survival data. Three actual microarray data sets for ovarian cancer were used. A. The cancer genome atlas (TCGA) data set [6] was used as the training data set; B. Tothill's data set [7] was used as a validation data set; and C. Bonome's data set [8] was used as a validation data set.

weighted by the mixing parameter α and L2-norm weighted by (1-α) had intermediate properties between L1-norm and L2-norm. When the mixing parameter α was larger, the contribution of L1-norm is larger, and when the mixing parameter α was smaller, the contribution of L2-norm was larger (see Figures S2, S3, S7, and S8).

The TPR that indicates the sensitivity for the penalized methods was larger than the TPR for the stepwise method when the sample size was larger than the number of candidate genes and the true value of coefficients was comparatively large. A larger contribution of L2-norm led to a higher TPR (see Figures S4 and S9). The PPVs for the penalized methods, other than the ridge method that uses all the candidates, were larger than for the stepwise method when the number of the candidate genes was larger than the sample size. A larger contribution of the L1-norm gave a larger value for the PPV (see Figures S5 and S10).

The statistical power of a validation set for the penalized methods, except the ridge method, in many cases was larger than the statistical power of the stepwise methods, particularly when $n < p$. Thus, when the number of candidate genes is larger than the sample size, a reasonably effective and reproducible narrowing-down of candidate genes may be possible using the penalized methods, including both L1-norm and L2-norm.

When the survival data included censored cases, the SDs of coefficients for each method increased and the TPRs, PPVs, and the statistical powers for each method decreased compared with the uncensored data (see Figures S6, S7, S8, S9, and S10). This means that

because of the increase in the number of censored cases, and thus a decrease of information about the events, the prediction abilities of the model decreased.

With increasing the correlation coefficients between each gene, the numbers of selected genes and the SDs of coefficients for each method increased. The TPR and the PPVs for each method decreased (see Figures S11 and S12). That means that due to the increase of the correlation coefficients between each gene, and thus the increase of the noisy information about the disease susceptibility genes followed by, the prediction abilities decrease. With increased correlation coefficients between each gene, the statistical powers of the validation sets seemed to increase for the penalized method; however, because this phenomenon was accompanied by a remarkable increase in the numbers of selected genes, the shrinkage effect was lost. That is, although the increase in statistical power looked good, it was actually not good because the number of selected genes increased and the aim of decreasing the number of genes was not achieved. This suggests that the increase of correlation coefficient is not good for these penalized methods. However, the rate of decrease of the TPRs for the elastic net method was smaller than for the lasso method, meaning that the penalized methods including both L1-norm and L2-norm may have a more narrowing down effect for the true genes with correlation compared with other penalized methods (see Figures S12).

In this study we analyzed actual DNA microarray expression survival data using a penalized Cox's proportional hazards model. To use the penalized Cox's proportional hazards model, some important points should be noted. One such point is the choice of the optimal value of the tuning parameter λ and the optimal mixing parameter α. The package for the two-dimensional cross-validation of λ and α has been reported previously [9]; however, this package supports only categorical and plain numeric data, so the survival data will not be passed through the function correctly.

Therefore, in the present study, the two-dimensional cross-validation was performed to choose the optimal λ and α using the pre-computed same fold vector in separate calls to cross-validate with different values of α. (i.e. two parameters are selected by the same fold with different values of α, then another two parameters are selected by another same fold with different values of α, and so on).The optimal value of λ was determined using the cross-validation method with different values of α. The optimal value of α was determined as follows: first, the candidate mixing parameter α was set to between 0.0 and 1.0 in intervals of 0.05; then, the optimal α was selected as the point where the graphs of the TPR, TNR, PPV, NPV, and statistical powers crossed each other. In this way, the actual microarray gene expression survival data were verified and their good reproducibility was confirmed.

The simulation experiments and verified actual data examples showed that each of the penalty methods reduced overfitting, thereby improving the precision of prognostic prediction. In particular, the statistical power of validation sets for the penalized methods including both the L1-norm and the L2-norm may be the largest. However, some bias may have been introduced by the penalty term; therefore, the estimate may not be the best unbiased estimate. In such cases, penalized methods have a comparatively high value as primary screening analyses tools for the identification of the leading disease susceptibility genes from a collection of candidate genes.

Conclusion

Our simulation programs for the Cox's proportional hazards model regularized by various penalties are very useful for planning DNA microarray studies or for evaluating the results of such studies.

Acknowledgements

NK is very grateful to Department of Medical Informatics and Department of Obstetrics and Gynecology in Niigata University Medical and Dental Hospital for funding this project.

References

1. Li J, Lenferink A, Deng Y, Collins C, Cui Q, et al. (2010) Identification of high-quality cancer prognostic markers and metastasis network module. Nat Communs 1: 34.
2. Tibshirani R (1997) The Lasso Method for Variable Selection in the Cox Model. Stat Med 16: 385–395.
3. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33: 1–22.
4. Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw 39: 1–13.
5. Friedman J, Hastie T, Tibshirani R (2014) Package 'glmnet' Lasso and Elastic-Net Regularized Generalized Linear Models. R package version 1.9: 8.
6. TCGA Data portal (2011) TCGA data set (Affymetrix HT-HG-U133A).
7. Tothill RW, Tinker AV, George J, Brown R, Fox SB, et al. (2008) Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res 14: 5198–5208.
8. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, et al. (2008) A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res 68: 5478–5486.
9. Kuhn M (2014) R 'caret' package version 6: 0-30.