RESEARCH ARTICLE

Survival Prognostic Factors of Male Breast Cancer in Southern Iran: a LASSO-Cox Regression Approach

Hadi Raeisi Shahraki[1], Alireza Salehi[2], Najaf Zare[3]*

Abstract

We used to LASSO-Cox method for determining prognostic factors of male breast cancer survival and showed the superiority of this method compared to Cox proportional hazard model in low sample size setting. In order to identify and estimate exactly the relative hazard of the most important factors effective for the survival duration of male breast cancer, the LASSO-Cox method has been used. Our data includes the information of male breast cancer patients in Fars province, south of Iran, from 1989 to 2008. Cox proportional hazard and LASSO-Cox models were fitted for 20 classified variables. To reduce the impact of missing data, the multiple imputation method was used 20 times through the Markov chain Mont Carlo method and the results were combined with Rubin's rules. In 50 patients, the age at diagnosis was 59.6 (SD=12.8) years with a minimum of 34 and maximum of 84 years and the mean of survival time was 62 months. Three, 5 and 10 year survival were 92%, 77% and 26%, respectively. Using the LASSO-Cox method led to eliminating 8 low effect variables and also decreased the standard error by 2.5 to 7 times. The relative efficiency of LASSO-Cox method compared with the Cox proportional hazard method was calculated as 22.39. The19 years follow of male breast cancer patients show that the age, having a history of alcohol use, nipple discharge, laterality, histological grade and duration of symptoms were the most important variables that have played an effective role in the patient's survival. In such situations, estimating the coefficients by LASSO-Cox method will be more efficient than the Cox's proportional hazard method.

Keywords: Cox proportional hazard - high dimension - LASSO - low sample size - male breast cancer survival

Asian Pac J Cancer Prev, **16 (15)**, 6773-6777

Introduction

Male breast cancer (MBC) is a rare disease with an incidence rate of less than 1% of the female breast cancer (FBC). In a large population-based study in northern European counties and Singapore, the world standardized incidence rates of breast cancer were 66.7 per 10^5 person-years in women and 0.40 per 10^5 person-years in men. Women were diagnosed earlier than men by about 8 years (Miao et al., 2011). In Iran, the incidence rate of FBC was about 148 per 10^5 (Zare et al., 2013) and the mean age was about 48 years (Zare et al., 2012, 2013) while in men the mean age was about 60 years (Salehi et al., 2011). Some demographic and clinico-pathologic prognostic factors such as age at diagnosis, tumor grade and lymph node status have been shown to be associated with overall survival (Miao et al., 2011; Salehi et al., 2011; Zare et al., 2012, 2013; Soliman et al., 2014).

Because of the rarity of this disease, the management of MBC patients is generalized from FBC which suffer from lack of evidence -based data to support this female to male extrapolation. The smallness of the existing sample size in MBS studies (Egypt, Turkish, Iran) resulting from the scarcity of male breast cancer on the one hand, and the high number of the independent variables that can potentially influence the patients' survival on the other hand, has been challenging and resulted in imprecise findings (Gui and Li, 2005).

In the studies where the dependency of survival time with regard to the independent variables is desirable, the Cox proportional hazard model is used to estimate the survival time. However, in the settings where the number of independent variables is high and sample size is low, analysis of the survival data is faced with a serious challenge. Problems such as multicollinearity, reduction in estimation precision, lack of a sparse model, and non-interpretability of the coefficients obtained from the Cox proportional hazard method have made this method an inefficient and invalid one in dealing with such data (Cox, 1972; Tibshirani, 1997). Even the numerous techniques of variable selection in regression which have also been generalized in Cox method aiming to keep all the variables in the model become inefficient (Gui and Li, 2005).

In 1997, in order to choose the most important variables under the Cox proportional hazards model by adding a penalized function to the estimation of the partial

*[1]Department of Biostatistics, [2]Research Center for Traditional Medicine and History of Medicine, [3]Department of Biostatistics, Infertility Research Center, Shiraz University of Medical Sciences, Shiraz, Iran *For correspondence: najafzare@sums.ac.ir*

maximum likelihood, Tibshirani introduced his method titled LASSO (Least Absolute Shrinkage and Selection Operator)-. Through placing a constraint on the absolute value of the regression coefficients, this penalized function causes many of the coefficients to get smaller and also some of the coefficients to become exactly zero. By omitting additional and redundant variables and making a brief bias in the model if necessary, LASSO-Cox method controls multicollinearity and is also simply applicable even in the settings where the number of variables is higher than the sample size (Tibshirani, 1996; Tibshirani, 1997).

In this research, we aimed to identify the most probable prognostic factors effective on survival of male breast cancer through the method of LASSO-Cox.

Materials and Methods

In this study, the data were obtained from the cancer registry of Vice-Chancellor for Health Affairs of Shiraz University of Medical Sciences and Shiraz hospitals during January 1, 1989 and January 1, 2008. During the study period, 63 histological proven MBCs were identified. The attainable probable prognostic factors were as follow: age at diagnosis, residence, history of alcohol use, nipple discharge, nipple ulceration, nipple retraction, skin fixation, skin redness, laterality, location of tumor, tumor size, axillary lymph node involvement, chest wall invasion, duration of symptoms, staging, and grading.

The information about the patients' survival was obtained from the Death Registry of Vice-Chancellor for Health Affairs of Shiraz University of Medical Sciences and telephone contacts were made to complete the information. After excluding the individuals for whom the survival time had not been recorded, the number of patients reduced from 63 to 50. Appropriate classification was done for all variables and also dummy variables were used to represent the data as zero and one.

A multiple Cox proportional hazards model was used to develop a predictive model of overall and disease-free survival, based on demographic and clinical covariates. The model has the following form $h(t|x)= h_0(t)\exp\{\beta^T X\}$, Where $X=(X_1, X_2, \ldots, X_p)$ are covariates, $h(t|x)$ is the hazard at time t, h0(t) is the unspecified baseline hazard function, and $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of regression coefficients (Cox, 1972).

Due to the noted shortcomings of the Cox proportional hazard model and the high correlation between some covariates, the variable selection was done by incorporating a LASSO (least absolute shrinkage and selection operator), or L1 penalty, on the regression coefficients $\beta_1, \ldots, \beta_p$. The LASSO penalizes the size of the parameter vector, β, so that unimportant variables (variables whose β coefficients are close to zero) are removed from the model. This results in a penalized log partial likelihood function of the form $l(\beta) - \sum_{j=1}^{p} \lambda|\beta_j|$, where $l(\beta)$ denotes the standard Cox log partial likelihood. The maximum likelihood estimates β are those which maximize this penalized likelihood. The parameter λ is the shrinkage parameter and determines the extent of variable selection, with larger values corresponding to a larger penalty and a greater number of variables removed. The optimal value for λ was determined using 5-fold cross-validation.

A multiple Cox proportional hazards model was used to develop a predictive model of overall and disease-free survival, based on demographic and clinical covariates. The model has the following form $h(t|x)= h_0(t)\exp\{\beta^T X\}$, where $X=(X_1, X_2, \ldots, X_p)$ are covariates, $h(t|x)$ is the hazard at time t, $h_0(t)$ is the unspecified baseline hazard function, and $\beta = (\beta_1, \ldots, \beta_p)$ is the vector of regression coefficients (Cox, 1972).

Cox proportional hazard model is expressed as $h(t|x)= h_0(t) \exp\{\beta^T X\}$, where $h_0(t)$ is a baseline hazard function, $X=(X_1, X_2, \ldots, X_p)$ is the vector of independent variables, and $\beta=(\beta_1, \beta_2, \ldots, \beta_p)^T$ is the vector of regression coefficients. Cox partial likelihood function is defined as follows:

$$PL(\beta)= \prod_{r\in D} \frac{\exp(\beta^T X_j)}{\{\sum_j \exp(\beta^T X_j)\}}$$

where D is the risk set of the events. The estimated coefficients of Cox's method is the value that maximize the above function(Cox, 1972).

$$\hat{\beta}_{Cox} = \text{argmax}\ \{PL(\beta)\}$$

LASSO that was first made in linear regression and then generalized to logistic and Cox regression, both estimates and selects the variables simultaneously by using a penalized function. LASSO-Cox regression coefficients are obtained through solving the equation:

$$\hat{\beta}_{\text{LASSO Cox}} = \text{argmax}\ \{PL(\beta)\} \quad \text{subject to } \sum_{j=1}^{p}|\beta_j| \le t$$

where t is a positive constant (Tibshirani, 1996; Tibshirani, 1997).

If you look at the variable selection procedure by LASSO method in Bayesian approach, this method uses in fact double exponential distribution (Laplace) as the prior distribution of regression coefficients:

$$f(\beta_j)=\frac{1}{2\tau}\exp\{-\frac{|\beta_j|}{\tau}\},$$

where $\tau = \frac{1}{\lambda}$

This approach increases the probabilities of zero points

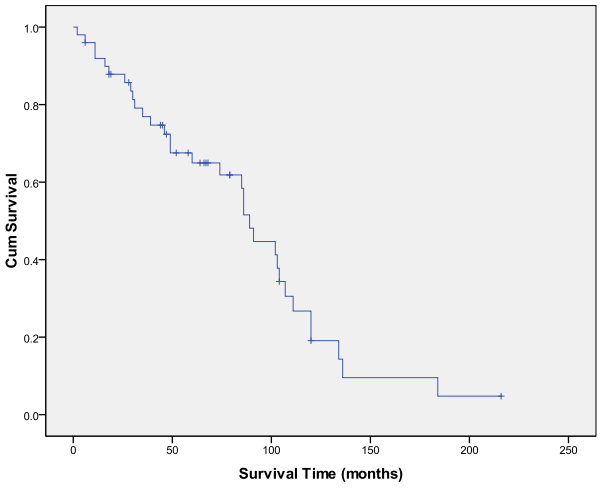

Figure 1. Survival Time of Patients in Month

and tails, and instead decreases the probabilities of the middle points (Tibshirani, 1996).

LASSO-Cox method coefficients can be calculated equivalently from the following equation:

$$\hat{\beta}_{\text{LASSO Cox}} = \text{argmax}\ \{PL(\beta) + \lambda \sum |\beta_j|\}$$

, where $PL(\beta)$ is Cox partial likelihood function, and λ is positive constant value called tuning parameter. Choosing the optimal λ is very important because the higher this value is, the more coefficients become zero, the model gets more sparse, and also the higher the interpretability will be.

In order to get the optimal value, a variety of cross validation methods like 5-fold, 10-fold and generalized cross-validation can be used. In K-fold cross-validation method, the data is divided into k equal subsets. In each

Table 1. Probable Prognostic Factors of Survival in Male Breast Cancer Patients by Cox and LASSO-Cox Methods

		Cox (n=50)		LASSO – Cox (n=50)	
	n (%)	Beta (SE)	Hazard	Beta (SE)	Hazard
Age					
<50	12 (19.0)	-	-	-	-
50 – 64	29 (46.0)	-3.6 (3.85)	0.03	-1.56(1.48)	0.21
>64	22 (35.0)	-0.84 (3.2)	0.43	0	1
Residence					
Fars	50 (88.3)	-	-	-	-
Other provinces	10 (16.7)	-1.03(3.25)	0.36	-0.37(0.77)	0.69
History of Alcohol Use					
No	48 (77.4)	-	-	-	-
Yes	14 (22.6)	-0.51(3.11)	0.6	-0.78(1.06)	0.46
Nipple Discharge					
No	53 (84.1)	-	-	-	-
Yes	10 (15.9)	-5.59 (5.41)	0	-1.27(1.34)	0.28
Nipple Ulceration					
No	55 (87.3)	-	-	-	-
Yes	8 (12.7)	2.59(4.29)	13.33	0	1
Nipple Retraction					
No	56 (88.9)	-	-	-	-
Yes	7 (11.1)	5.62(5.33)	275.89	0.25(0.73)	1.29
Skin Fixation					
No	53 (84.1)	-	-	-	-
Yes	10 (15.9)	0.8 (3.6)	2.23	0	1
Skin Redness					
No	56 (88.9)	-	-	-	-
Yes	7 (11.1)	-4.15(6.41)	0.02	0	1
Laterality					
Left	30 (49.2)	-	-	-	-
Right	31 (50.8)	-3.13(3.34)	0.04	-1.16(1.07)	0.31
Location of Tumor					
Retroareolar	47 (81.0)	-	-	-	-
Other quadrant	11 (19.0)	0.61(3.87)	1.84	0	1
Chest Wall Invasion					
No	54 (90.0)	-	-	-	-
Yes	6 (10.0)	0.31(3.9)	1.36	0	1
Axillary Lymph Node Involvement					
No	32 (54.2)	-	-	-	-
Yes	27 (45.8)	0.54 (2.81)	1.72	0	1
Stage					
Stage 1	6 (11.5)	-	-	-	-
Stage 2	39 (75.0)	-0.89(4.28)	0.41	0.24(1.22)	1.27
Stage 3	7 (13.5)	0.25(5.05)	1.28	0	1
Tumor size					
<2 cm	9 (18.8)	-	-	-	-
2 – 4.9 cm	27 (56.2)	1.86 (4.28)	6.42	0.09(0.94)	1.1
>4.9 cm	12 (15.0)	2.21(5.22)	9.12	0.26(1.28)	1.3
Histological Grade					
Grade 1	19 (41.3)	-	-	-	-
Grade 2 or 3	27 (58.7)	2.91(3.46)	18.36	1(1.35)	2.71
Duration of Symptoms (month)					
Symptom≤6	20 (44.4)	-	-	-	-
6< Symptom ≤12	14 (31.1)	0.95(3.82)	2.59	0.15(1.02)	1.16
12< Symptom	11 (24.5)	-1.35(3.56)	0.26	-1.11(1.26)	0.33

time, one of these subsets is considered as the validation data and the error is calculated by using the obtained estimations for the other subsets (training data). This is repeated so much that each subset is used just once as the validation data. The error mean is considered as c(λ)in all the repetitions, and the value of λ for which c(λ) becomes minimum is chosen as the optimal tuning parameter (Hastie et al., 2009; Goeman, 2010). In generalized cross validation method, the value of λ which minimize the following value is chosen as the optimal λ:

$$GCV=\frac{1}{n}\sum_{i}\frac{(y_i-\hat{y}_i)^2}{1-tr(H)/n}$$

Where $H=(X^T X)^{-1} X^T Y$ (Craven and Wahba, 1978).

In this research λ values for each data set were obtained separately through generalized cross validation method. Because of having missing information in some variables, multiple imputation was used 20 times in Markov Chain Mont Carlo method, and suitable values were imputed. The advantage of this method is that it allows the researchers to use most of the existing information without violating the validity of the results (Goeman, 2010).

For each one of the 20 obtained sets of data, Cox proportional hazard and LASSO-Cox model was fitted independently and coefficients were reported. Standard error of LASSO-Cox coefficients were obtained through bootstrap method with 1000 time repetitions.

For the variables whose coefficients become zero in 50 or more percentages of the times, zero coefficients were reported and for the other variables the mean of the non-zero coefficients in 20 data sets was reported as the coefficients of that variable. In the remaining variables, the Robin's formula was used in order to report the standard error of the non-zero regression coefficients as follows:

$$SE(Q_i)=\sqrt{(Var\ (Q_i)}=\sqrt{(U_i+(1+\frac{1}{k_i})B_i}$$

In the above formula Q_i, $\overline{U}$, B and k are the ith regression coefficients, the mean variance of the non-zero coefficients, sample variance of the non-zero coefficients and the number of non-zero regression coefficients, respectively (Rubin, 1977; Rubin, 2009; Goeman, 2010).

The relative efficiency of LASSO-Cox method versus Cox proportional hazard method was calculated by the following formula proposed by Casella and Berger (Casella and Berger, 1990):

$$\text{Relative efficiency} = \frac{\text{(sum of sequare of coefficients standard error obtained from Cox proportional hazad method)}}{\text{(sum of sequare of coefficients standard error obtained from LASSO-Cox method)}}$$

Data analysis was done through SPSS 16.0 and R-3.0.1.Software.

Results

Of the 63 patients, 38 were alive during the study, 18 died and 7 had missing history of death. All available data were used in descriptive reports, but for analytical purpose, we used the complete or imputed information of 50 patients.

In 50 patients, the age at diagnosis was 59.6 (SD=12.8) years with a minimum of 34 and maximum of 84 years and the mean of survival time was 62 months. 18 patients

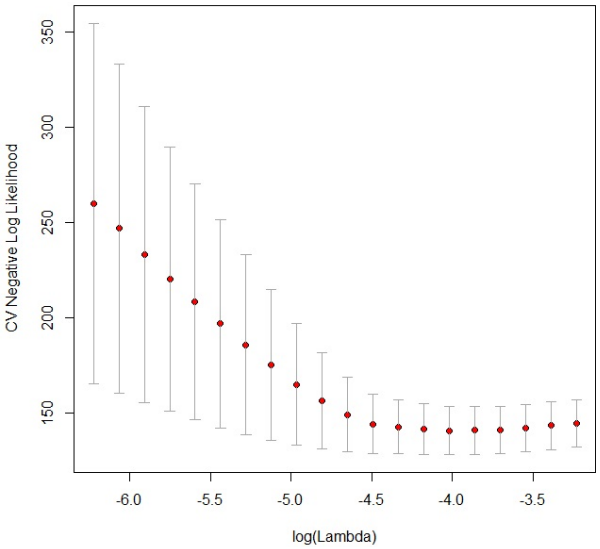

Figure 2. The Generalized Cross Validation Negative Log Likelihood. Plotted Against the Logarithm of Lambda

(36%) had died and the other (64%) had been censored. Three, 5 and 10 year survival were 92%, 77% and 26%, respectively (Figure1).

The negative logarithm of likelihood function reached its minimum, i.e 140.78 in log λ=-4.02 resulting the optimal λvalue has equaled to 0.018. The same procedure was implemented for the other data sets. The variables under the study, manner of classification, and the results of fitting the two models are shown in Table 1. The results indicate that using LASSO-Cox method led to getting zero 8 variable with low effects, while Cox's method has retained all variables in the model. Besides, standard errors of the regression coefficients in Cox proportional hazard method are 2.5 to 7 as high as their correspondent ones in LASSO-Cox method (Table 1).

In order to do a total comparison between the two methods, the relative efficiency of LASSO-Cox method compared with the Cox proportional hazard method was calculated as 22.39, which is indicative of the of LASSO-Cox method being 22 times more efficient.

Discussion

The 19 year survival of the male breast cancer in Fars province showed that age, a history of alcohol use, nipple discharge, laterality, histological grade and duration of symptoms were the most important variables that can play an effective role in the patient's survival. By omitting the variables being low in effectiveness, LASSO-Cox method could estimate the most important variables much more exactly and efficiently than the Cox proportional hazard method.

Large values of standard errors in Cox proportional hazard method can represent the presence of multicollinearity among the variables which has caused the lack of stability and high variability of regression coefficients from one data set to another. In such circumstances, even the regression coefficients represented by Cox proportional hazard method will not

be reliable.

The coefficients obtained from LASSO-Cox method represent an increase in grade and tumor size as the variables that can cause a decrease in survival time; these results are in the same line with those of most studies on this ground, while they do not agree with a few other studies (Kuroi and Toi, 2003; Fentiman et al., 2006; Salehi et al., 2011).

In this study, nipple discharge and alcohol consumption were introduced as two factors affecting the increase in patient's survival. Regarding the few studies conducted on the effects of these two variables on the survival time, decisive remarks are hard to be made.

As a limitation of this study, we can refer to the lack of considering such variables as marital status, metastasis and undergoing a variety of treatments. Our study is the first one in Islamic Republic of Iran that measures the simultaneous effect of several variables on the survival of cancer patients. While this study has used all the information obtained from Shiraz hospitals as the center of south of Iran, planning multicenter studies in different settings on larger sample sizes seems necessary.

Acknowledgements

The authors would like to thank Dr. Nasrin Shokrpour at Center for Development of Clinical Research of Nemazee Hospital for editorial assistance.

References

Casella G, Berger RL (1990). Statistical inference, Duxbury Press Belmont, CA.

Cox DR (1972). Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 187-220.

Craven P, Wahba G (1978). Smoothing noisy data with spline functions. *Numer Math*, **31**, 377-403.

Fentiman IS, Fourquet A, Hortobagyi GN (2006). Male breast cancer. *Lancet*, **367**, 595-604.

Goeman JJ (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical J*, **52**, 70-84.

Gui J, Li H (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001-8.

Kuroi K, Toi M (2003). Male breast cancer]. Gan to kagaku ryoho. *Cancer Chemotherapy*, **30**, 599.

Miao H, Verkooijen HM, Chia K-S, et al (2011). Incidence and outcome of male breast cancer: an international population-based study. *J Clin Oncol*, **29**, 4381-6.

Rubin DB (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J Am Stat Associat*, **72**, 538-43.

Salehi A, Zeraati H, Mohammad K, et al (2011). Survival of male breast cancer in Fars, South of Iran. *Iranian Red Crescent Med J*, **13**, 99.

Soliman AA, Denewer AT, El-Sadda W, et al (2014). A retrospective analysis of survival and prognostic factors of male breast cancer from a single center. *BMC cancer*, **14**, 227.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-88.

Tibshirani R (1997). The lasso method for variable selection in the cox model. *Stat Med*, **16**, 385-95.

Zare N, Doostfatemeh M, Rezaianzadeh A (2012). Modeling of breast cancer prognostic factors using a parametric log-logistic model in fars province, Southern Iran. *Asian Pac J Cancer Prev*, **13**, 1533-7.

Zare N, Haem E, Lankarani KB, et al (2013). Breast cancer risk factors in a defined population: weighted logistic regression approach for rare events. *J Breast Cancer*, **16**, 214-9.