L1 Penalized Cox Regression to Characterize Cardiovascular Events in Hypertensive Patients

Rafael García-Carretero[1,2], Óscar Barquero-Pérez[2], Inmaculada Mora-Jiménez[2], Cristina Soguero-Ruiz[2], Rebeca Goya-Esteban[2], Antonio G. Marques[2], Javier Ramos-López[2]

[1] Móstoles University Hospital, Móstoles (Madrid), Spain
[2] Rey Juan Carlos University, Fuenlabrada (Madrid), Spain

Abstract

Reliable cardiovascular risk stratification of hypertensive patients is essential to provide appropriate clinical management. However, traditional statistical approaches may ignore important information from datasets and overlook possible interactions among covariates when dealing with high-dimensional data. Regularization techniques, such as least absolute shrinkage and selection operator (LASSO), may improve the prediction accuracy and interpretability of regression models. Our aim is to identify the most relevant features to predict cardiovascular (CV) events in hypertensive patients by using the L^1-penalized Cox regression. We use clinical records of 1664 patients of the Móstoles University Hospital whose CV status was determined by clinical variables and biomarkers including body index mass, blood pressure, cholesterol, albumin/creatinine ratio, and kidney function. By monotonically tuning its regularization parameter, the LASSO approach put forth is able to identify the most predictive features, with "cystanin C-based glomerular filtration" being the single most relevant predictor for CV events.

1. Introduction

Essential hypertension causes long-term adverse effects and is one of the most important risk factors for cardiovascular (CV) disease, including heart failure, stroke, myocardial infarction, and chronic kidney disease [1, 2]. Appropriate clinical management of hypertensive patients requires reliable CV risk stratification. A key step to carry out such a stratification is the identification of the features (clinical variables and biomarkers) with the strongest impact in the probability of patients to suffer from a CV event. Classical feature selection methods using univariate analysis may be impractical or inefficient, especially when dealing with many features or collinearity [3], since they may overlook hidden relationships among features and clinical outcomes such as mortality and CV events [4]. Overfitting may be another important issue, which is defined as the effect of describing a random effect rather than the real underlying relationship among features. Regularization is a common technique to avoid overfitting in linear regression models [4]. The least absolute shrinkage and selection operator (LASSO) is a rigorous way to address these two problems jointly. By augmenting the fitting regression term with a L^1-norm cost, LASSO simultaneously performs regularization (leading to a more generalizable model) and feature selection (enforcing sparsity in the final solution). This approach has been also applied in genome-wide association analysis [5] and when trying to identify prognostic factors with high-dimensional data such as radiological features of PET images [6] or environmental enteropathy biomarkers [7]. In these situations, traditional statistical methods for feature selection may be tedious or inefficient due to the amount of covariates and the non-obvious correlation among them.

The focus of this paper is on the identification of the features (variables and biomarkers) that have the strongest impact on the evolution of hypertensive patients. The ultimate goal would be to use such features as prognostic factors to predict the risk of a patient from suffering a CV event. Our technical contribution is the application of an L^1-penalized Cox regression approach to the dataset at hand. The Cox regression offers a simple but efficient way to model the impact of the different features on the probability of the CV event [8], while the L^1-penalization serves to select the most predictive features and avoid overfitting. Our approach uses data from more than 1664 hypertensive patients from the Móstoles University Hospital.

2. LASSO model

LASSO is a regularization linear regression method that shrinks coefficients towards zero, promoting sparse solutions (several coefficients equal to 0) and, therefore, improving the model interpretability. Given a linear regression model:

$$\hat{y} = \boldsymbol{\beta}^T \boldsymbol{x} \tag{1}$$

 ISSN: 2325-887X DOI:10.22489/CinC.2017.253-209

Table 1. Baseline features of our hypertensive cohort.

Total Patients	1,664
Age (years)	56.8 ± 13.3
BMI	31.6 ± 5.6
Systolic BP (mmHg)	140.7 ± 12.4
Diastolic BP (mmHg)	79.3 ± 8.4
LDL cholesterol (mg/dL)	134.6 ± 32.6
HDL cholesterol (mg/dL)	64.0 ± 17.3
Tryglicerides (mg/dL)	154 ± 102
CRP (mg/dL)	6.1 ± 1.6
HbA1c (%)	5.9 ± 1.0
Albumin/creatinine ratio (mg/g)	16.1 ± 46.6
eGFRcreat (mL/min/1.73m^2)	92.6 ± 20.5
eGFRcyst (mL/min/1.73m^2)	102.2 ± 23.1

Data are reported as percentages or median (± interquartile range). BMI: body index mass. BP: blood pressure. CRP: C-reactive protein. HbA1c: glycated hemoglobin. eGFRcreat: estimate glomerular filtrate rate from serum creatinine. eGFRcyst: estimate glomerular filtrate rate from serum cystatin C. Both eGFRcreat and eGFRcyst were computed using CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) equations.

the goal is to estimate a response variable y, by a linear combination of explicative features (covariates) in column vector $\boldsymbol{x}$ using a set of coefficients, $\boldsymbol{\beta}$, which have to be estimated.

In LASSO, the regression coefficients $\boldsymbol{\beta}$ are estimated minimizing the following objective function:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{2}$$

where $\boldsymbol{y}$ is a column vector with the response variable for every individual, $\boldsymbol{X}$ is a matrix where features are in columns, and $\|.\|_p$ denotes the L^p-norm of a vector. The presence of the L^1-norm promotes solutions where $\boldsymbol{\beta}$ is sparse, with the regularization parameter λ controlling the particular number of entries of $\boldsymbol{\beta}$ that are zero [9, 10].

LASSO allows to perform a regularization path, in which profiles of the lasso coefficients are provided as the regularization parameter λ changes [10].

3. LASSO for Cox regression models

Survival analysis deals with the statistical modeling of the time when a (death) event takes place. Cox regression is a simple but effective way to perform survival analysis. Specifically, under this approach, the instantaneous probability of death (event) at time t, given survival (no event) up till t, i.e. *hazard function*, is modeled by:

$$h(t, \boldsymbol{x}) = h_0(t)e^{\boldsymbol{\beta}^T\boldsymbol{x}} \tag{3}$$

where $h_0(t)$ is the hazard with $\boldsymbol{x} = \boldsymbol{0}$, i.e. baseline hazard. The value of the the i-th coefficient in $\boldsymbol{\beta}$ models the importance of the i-th feature in $\boldsymbol{x}$ in causing the event.

The coefficients $\boldsymbol{\beta}$ in a Cox regression with an L^1-norm penalization are estimated as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(-\sum_i \log\left[\frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}^T\boldsymbol{x}_j}}\right] + \lambda\|\boldsymbol{\beta}\|_1 \right) \tag{4}$$

where for each $i = 1, \ldots, N$, R_i is the set of individuals of the study who are alive at time t_i.

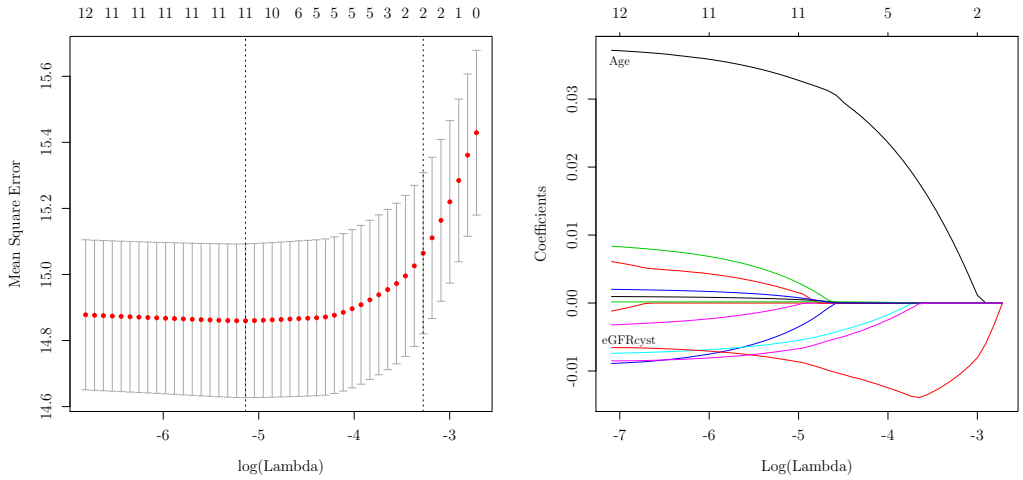
The first term in (4) corresponds to the log of the partial likelihood and serves as counterpart of the least squares fitting cost for the linear regression in (2). As in the regular LASSO, the second term promotes sparsity on $\boldsymbol{\beta}$.The regularization parameter λ is commonly chosen using k-fold cross-validation, where k is usually between 5 and 10 [11].

4. Description of database and statistical analyses

Hypertensive patient records were collected from the Hypertension Unit of Móstoles University Hospital between 2006 and 2016. Patients with prevalent CV disease were excluded. The cohort included 1664 patients, out of 3473, and the follow-up was 11.2 years (median 4.6 years). Around 34.8% of patients in the database were diabetic, and 51% were women. Demographic, clinical and biochemical data were collected, and kidney function was calculated using the CKD-EPI equations [12]. Table 1 shows the baseline value for the features used in this work. Baseline creatinine and cystatin-C were 0.8±0.2 mg/dL (median± interquartile range). The CV events were myocardial infarction (37), heart failure (27), stroke (46), and death (47).

The Cox regression model with L^1-norm penalization (LASSO) was performed to: (1) feature selection; (2) explore the coefficients path, and (3) build a predictive model for time-to-event data. In this work, we compared features identified as significant using L^1-penalized and classical Cox regression models. Several features were correlated, such as age, C-reactive protein and kidney function calculated by both creatinine-based eGFR (eGFRcreat) and cystatin C-based eGFR (eGFRcyst). We performed a two-stage analysis, since our aim was to compare how both survival methods (L^1-penalized and classical) addressed the collinearity issue.

In the first stage, we selected features using two approaches. The first approach (A1) considers the correlation coefficient to check the statistical relationship between features and outcome. Those features with a statistical association were suitable for a univariate Cox-regression. The second approach (A2) performs the LASSO method to reduce the number of significant features. In the second stage, we constructed two multivariate survival models through a survival analysis with features selected by both approaches.



(a) (b)

Figure 1. Influence of the regularization parameter λ. Numbers at the top of the plots represent the number of features of the model as λ changes. (a) Confidence intervals for the Mean Squared Error (MSE) when performing 10-fold cross validation. Vertical dotted line on the left marks the λ value for the minimum MSE. Vertical dotted line on the right indicates the λ chosen according to the parsimonious model (MSE is within one standard deviation of the minimum MSE); (b) Path of the coefficients (*lassopath*) for every feature when using the penalized Cox-regression under L^1-norm.

5. Results

Some features in Table 1 were strongly correlated, such as age, CRP, and kidney function calculated by both eGFR-creat and eGFRcyst.

On the one hand, only 5 features (age, diastolic BP, LDL-cholesterol, eGFRcreat, and eGFRcyst) were considered for the multivariate Cox-regression analysis performed by A1. That is, A1 excluded BMI, systolic BP, HDL-chol, triglycerides, HbA1c, CRP and albumin/creatinine ratio because corresponding correlation coefficients with the outcome were low.

On the other hand, the L^1-penalized Cox regression was performed to automatically discard irrelevant features. The regularization parameter λ was selected using 10-fold cross-validation and the parsimonious model, see Figure 1(a). With this value of λ , approach A2 just identified features age and eGFRcyst as predictors (see Figure 1(b)). The LASSO regression coefficient related to age was positive ($\beta_{age} = 0.01$), indicating that the older the patient is, the greater the CV risk. However, the regression coefficient related to eGFRcyst was negative ($\beta_{eGFRcyst} = -0.01$), indicating that it is a protective factor, so the greater its value, the lower the CV risk. This finding is biologically plausible: the higher the eGFRcyst, the healthier the patient is. Note also from Fig. 1(b) how the increase in λ is related to a reduction in the number of selected features, being eGFRcyst the single most relevant predictor for CV events (last coefficient to be set to zero).

To further evaluate the prognostic features for our primary outcome, we performed the multivariate Cox analyses for both approaches. When calculating hazard ratios using the features in A1, only age (HR 1.03, 95% CI 1.01-1.05, p=0.0001), eGFRcyst (HR 1.8, 95% CI 1.05-3.09, p=0.03), and LDL-cholesterol (HR 0.99, 95% CI 0.98-0.99, p=0.009) were statistically significant. On the other hand, when hazard ratios were calculated according to the model provided by A2, both age (HR 1.05, 95% CI 1.03-1.07, p <0.0001) and eGFRcyst (HR 2.55, 95% CI 1.62-4.01, p <0.0001) were significant.

6. Conclusions

The aim of this work was to assess the applicability of a regularized method, the L^1-penalized Cox regression, for predicting CV events on a population of hypertensive patients with correlated features. Since it is possible that only a small number of features are truly informative, we were interested in a method that implicitly drops correlated, non-relevant features, and confounding factors.

By discarding the contribution of less important covariates, the L^1-penalized Cox regression produced a parsimonious and biologically plausible model improving interpretability when compared with traditional statistical methods. It identified not only age as a prognostic factor, but also kidney function based on cystatin-C, a biomarker with increasing interest due to its relevance as a stronger predictor of CV disease than creatinine itself, what is in

line with other studies [13, 14].

In comparison with the classical approach, figures (hazard ratios, confidence interval, and p-values) provided by the L^1-penalized model are quite similar to those provided by the classical approach. The main difference is that the penalized method was more parsimonious as it discards the LDL-cholesterol feature. From a clinical viewpoint, this is an interesting result since patients in the database are overweight, with high LDL-cholesterol values, and so including the last feature is not going to provide more information about the CV risk. The simpler the model, the easier is its application and interpretation.

Acknowledgements

This work has been partly funded by Research Projects TEC2016-75361-R, and TEC2016-75161-C2-1-R from the Spanish Government, and Research Project DTS17/00158 from Instituto Carlos III (Spain). The authors declare that they have no conflict of interest.

References

[1] Grover SA, Hemmelgarn B, Joseph L, Milot A, Tremblay G. The role of global risk assessment in hypertension therapy. Canadian Journal of Cardiology 2006;22(7):606–613.

[2] Viazzi F, Leoncini G, Pontremoli R. Global cardiovascular risk assessment in the management of primary hypertension: the role of the kidney. International journal of hypertension 2013;2013.

[3] Schisterman EF, Perkins NJ, Mumford SL, Ahrens KA, Mitchell EM. Collinearity and causal diagrams–a lesson on the importance of model specification. Epidemiology Cambridge Mass 2017;28(1):47.

[4] Vasquez MM, Hu C, Roe DJ, Chen Z, Halonen M, Guerra S. Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. BMC medical research methodology 2016;16(1):154.

[5] Papachristou C, Ober C, Abney M. A lasso penalized regression approach for genome-wide association analyses using related individuals: application to the genetic analysis workshop 19 simulated data. In BMC proceedings, volume 10. BioMed Central, 2016; 53.

[6] Yue Y, Osipov A, Fraass B, Sandler H, Zhang X, Nissen N, Hendifar A, Tuli R. Identifying prognostic intratumor heterogeneity using pre-and post-radiotherapy 18f-fdg pet images for pancreatic cancer patients. Journal of gastrointestinal oncology 2017;8(1):127.

[7] Lu M, Zhou J, Naylor C, Kirkpatrick BD, Haque R, Petri WA, Ma JZ. Application of penalized linear regression methods to the selection of environmental enteropathy biomarkers. Biomarker research 2017;5(1):9.

[8] Kleinbaum DG. Survival analysis, a self-learning text. Biometrical Journal 1998;40(1):107–108.

[9] Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society Series B Statistical Methodology 2007;69(4):659–677.

[10] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B Methodological 1996;267–288.

[11] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. CRC press, 2015.

[12] Inker LA, Schmid CH, Tighiouart H, Eckfeldt JH, Feldman HI, Greene T, Kusek JW, Manzi J, Van Lente F, Zhang YL, et al. Estimating glomerular filtration rate from serum creatinine and cystatin c. New England Journal of Medicine 2012;367(1):20–29.

[13] Peralta CA, Katz R, Sarnak MJ, Ix J, Fried LF, De Boer I, Palmas W, Siscovick D, Levey AS, Shlipak MG. Cystatin c identifies chronic kidney disease patients at higher risk for complications. Journal of the American Society of Nephrology 2011;22(1):147–155.

[14] Shlipak MG, Matsushita K, Ärnlöv J, Inker LA, Katz R, Polkinghorne KR, Rothenbacher D, Sarnak MJ, Astor BC, Coresh J, et al. Cystatin c versus creatinine in determining risk based on kidney function. New England Journal of Medicine 2013;369(10):932–943.

Address for correspondence:

Óscar Barquero Pérez
Department of Signal Theory and Communications
Rey Juan Carlos University
Camino del Molino s/n 28943 - Fuenlabrada (Madrid)
oscar.barquero@urjc.es