MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Logistic Regression Model: An Assessment of Variability of Predictions

Isabelle Colombet, Marie-Christine Jaulent, Patrice Degoulet, Gilles Chatellier

SPIM, Faculté de Médecine, 15 rue de l'école de Médecine, 75005 Paris, France
colombet@hegp.bhdc.jussieu.fr

Abstract

Risk prediction models available for cardiovascular prevention are statistical or based on machine learning methods. This paper investigates whether the logistic regression method can be considered as reference for validation of other methods. In order to test the stability of the predictions using this method, we performed two types of analyses on 50 random training and test samples drawn from the same database. In first analyses three models were obtained by forced entry of different sets of four variables. In second analyses, models were built with increasing number of predictive variables. The predictive performance was assessed by the area under the ROC curve. Although across-samples variability is low for a given model, it is large enough to lead to wrong conclusions when comparing different prediction methods. We also suggest that a low events-per-variable ratio alters the stability of a model's coefficients but does not affect the variability of prediction performance.
Keywords: *risk model, logistic regression, stability of prediction, predictive performance*

Introduction

Several authors showed the difficulty for physicians to subjectively estimate the cardiovascular risk of their patients [1]. Moreover, it has been suggested that using an estimate of global cardiovascular risk could be more relevant to guide decisions, than using binary representation (presence or absence) of risk factors data [2]. In other medical domains, the access for the practitioners to a numerical expression of risk does modify their behavior [3]. However, the practical use of a quantitative estimate of risk for an individual who does not belong to the original study population raises the issues of *accuracy*, *precision* and *reproducibility* of the risk estimate (i.e. *variability* of the model). *Accuracy* is summarized by both the discriminative performance of the model and its calibration to the real risk *Precision* is usually represented by variance or confidence interval of the risk estimate, within the population from which the model was inferred. *Reproducibility* can be assessed by estimating *accuracy* and *precision* of the model to some data that were not used to build it [4]. Data splitting is often proposed to appreciate to which extent the prediction performance is overestimated when measured on training data relative to test data [5]. In addition to the estimate of this bias, methods such as cross validation and bootstrapping allow to appreciate the reproducibility of the prediction performance across different samples of the same data [6 7]. Another approach can be to control for conditions under which the reproducibility of this performance would be improved. If these conditions are met, a simple data splitting may be valid to ensure a stable measure of predictive performance of a model in training as well as in test data.

In cardiovascular prevention, most available models used to predict cardiovascular risk are obtained by using statistical methods [8 9]. Machine learning methods are more and more explored and evaluated for risk prediction purposes in medical domains [6 10 11]. They are usually compared with algebraic models such as logistic regression or Cox regression models. However, how a logistic regression model can be stable enough to be considered as a reference to evaluate other methods remains unclear [12]. This paper aims at exploring whether or not multiple logistic regression models are precise and reproducible prediction tools. As any attempt to qualify a modeling method is highly dependent of the nature of data used for exploration, we choose to use real data from cardiovascular domain and we shall be aware that any conclusions drawn from this work pertain to these particular data.

Background

Several potential causes of instability have been identified and described by biomedical statisticians [12 13]. We shall not discuss the failure to comply with the important modeling hypotheses that are common to all general linear modeling methods, namely the linearity assumption, the distributional assumption and the hypothesis of independence of the covariates (referred to as "independent variables"). Other important causes of instability relate to the choice of predictive characteristics and to the number of

outcome events available in the data to fit the model (which can be referred to as the ratio of the number of outcome events in training data divided by the number of predictive variables in the model: Events Per Variable ratio, EPV). Including too many predictive variables in a regression model may improve the predictive performance on training data, but may alter it on validation data, reflecting a problem of overfitting. Peduzzi et al. report a simulation study on how the EPV ratio affects the variability of the coefficients in logistic regression analysis [14]. They decrease the number of events per variable in simulated data by selecting samples with less events. Below ten events per variable, they showed a high variability of model quantification (*i.e.* regression coefficients). Would this variability be the same if measured on the actual discriminative performance of the model as assessed by ROC curve for example ?

The objective of this work is to study the effect of several conditions relative to the data, which determine the variability and reproducibility a logistic regression model for risk prediction. We explored the variability of predictive performance of models built from different random samples of the same data. We also hypothesized that accuracy, precision and reproducibility can be influenced by 1) the nature of predictive variables and 2) the number of predictive variables related to the number of outcome events.

Materials and Methods

Database and sampling methods

The reproducibility of predictive performance of different models was examined on one hundred random samples of the same database.

The database was provided by the Montreal Heart Institute and consists in 376 coronary stenoses described by 17 morphological attributes, observed in 84 patients known to have presented a myocardial infarction. For each patient, the stenosis that induced the myocardial infarction was identified. The database therefore comprises 84 stenoses that were involved in an infarction process (culprit stenoses) and 292 control lesions. Records with missing data were excluded (21 stenoses, while considering all independent variables). Baseline characteristics of stenoses are presented in Table 1.

Different multiple logistic regression models were fitted to the data to predict the probability of myocardial infarction for a given stenosis according to its morphological attributes.

Fifty samples of 250 stenoses were randomly selected from the original data and used to fit the models (training samples). The remaining stenoses (105 to 126, depending of missing data exclusion) were used to test the performance of the models (test samples). Additionally, fifty other training and test samples were selected according to the same procedure but stratified on the outcome event, in order to get a fixed number of culprit lesions (i.e. 56 outcome events).

Table 1: Summary characteristics in culprit and control lesions

Morphological characteristics	Culprit lesions (n=84)	Control lesions (n=292)
Qualitative variables (frequencies, %)		
Vessel (V): Circonflex	23	31
Right coronary	39	34
Left ventricular	38	34
Main coronary	0	1
Position (Po) Distal	19	33
Medial	23	22
Proximal	58	45
Ostium (Os) non ostial	94	92
Ostial	6	8
Calcification (Ca) Massive	4	2
Moderate	7	6
Absent	89	92
Tortuous character (Tc)		
Important	6	7
Minimal	69	67
Moderate	25	27
Outlines (Ol) Irregular	33	20
Smooth	60	78
Subocclusive	0	0.3
Ulcerated	7	2
Thrombus (Th) Certain	0	0.7
Ambiguous	11	3
Absent	88	97
Probable	1	0
Territory* (Te)		
Large	10	7
Medium	24	21
Non appropriate	46	49
Occluded	1	0
Small	19	23
Quantitative variables (means)		
Cyclic flexion (Cf)	10.5	12.1
Reference diameter (R)	2.9	2.7
Degree (D)	45.3	40.1
Length (L)	10.3	8.7
Symmetry (S)	0.6	0.5
Plaque area (Pa)	6.7	5
Inflow angle (If)	-14.7	-15.3
Outflow angle (Of)	15.2	12.7
Diastolic angle (Da)	150.6	149.9

Analyses

Two series of analyses were performed with each of the 100 training-test paired samples:

1) Three models were fitted to the training sample by forced entry of three different sets of four predictive variables (models 1a-1c). Each three sets of variables, described in Table 2, were chosen according to univariate analyses: in model 1a, all four variables are highly predictive and in model 1b and 1c, at least two variables are significantly predictive.

2) Eight models were fitted by forced entry of 4, 6, 8, 10, 12, and 14, 16 and 17 independent variables (models 2a-

2h), so that the number of events per variable in the training-sample gradually decreases from 14 (56 events/4 variables) to 3.2 (56 events/17 variables).

Table 2: Descriptive statistics of distribution of the areas under ROC curves for model 1a (including four variables: Position, Outlines, Degree, Symetry) in 50 stratified samples

	AUC in training samples	AUC in test samples
Mean	**0,7227**	**0,6734**
SD	**0,022**	**0,048**
median	**0,7269**	**0,6700**
1st quartile	**0,7042**	**0,6449**
3rd quartile	**0,7361**	**0,7071**
minimum	**0,6766**	**0,5562**
maximum	**0,7689**	**0,7820**

For each of these analyses, the first step of model fitting on the training-samples was followed by a second step of validation of the models on both training and test-samples. For a given model, validation consisted in computing the area under the curve (AUC) for each of the 100 samples and in summarizing them using the mean and standard deviation (within-model variability). We then assessed the internal variability due to the different modeling conditions by computing the overall standard deviation of the means obtained for all models (between-model internal variability). Reproducibility was assessed by comparison of means of AUCs and their standard deviation obtained in training samples with those obtained in test samples (ratio of the mean in training by the mean in test samples). ROC curves were compared using a non parametric approach based on the Mann-Whitney-U-statistic [15]. The R statistical software was used for analyses [16].

Results

Within-model variability due to sample fluctuations

For the model 1a, a variance of 0.022 across the 50 outcome-stratified training-samples reflected a low variability of the areas under the curve (AUCs) (see Table 3). However, these AUCs range from 0.68 to 0.77, the inter-quartile range being 0.70-0.74. When considering the models fitted on non-stratified training samples, ranges are larger (min-max: 0.66-0.78 and inter-quartile range: 0.70-0.77).

Between-model variability according to nature of variables

Distribution of the areas under the curve (AUCs) and their mean over the 50 stratified training-samples did not differ widely across the first three models 1a-1c (see Table 2). Figure 1 represents the ROC curves obtained with the three models for one stratified training sample: differences between the AUCs of the three models are respectively 0.042 (95% confidence interval: -0.0345; 0.042), 0.014 (95%CI: -0.014; 0.042) and -0.028 (95%CI: 0.1104; 0.0545) for models 1a-1b, 1a-1c and 1b-1c . The same differences were not significant either considering the AUCs on the corresponding test samples. The overall standard deviation of the three means (0.024) reflects a low effect of the nature of variables on internal variability.

Table 3: Summary of areas under the ROC curve in 50 training and test stratified samples, according to the nature and to the number of variables

Type of model	AUC training		AUC test	
	mean	SD	mean	SD
Model 1a* : Po,Ol,D,S	**0.72**	**0.022**	**0.67**	**0.048**
Model 1b : Ca,Ol,D,Pa	**0.68**	**0.024**	**0.64**	**0.054**
Model 1c : Tc,Ol,L,S	**0.72**	**0.021**	**0.68**	**0.046**
Mean of the 3 means (SD)	**0,71 (0,024)**		**0.66 (0.021)**	
Model 2a : 4 (14)#	**0.73**	**0,024**	**0.70**	**0,054**
Model 2b : 6 (9.3)	**0.66**	**0,023**	**0.63**	**0,057**
Model 2c : 8 (7)	**0.73**	**0,024**	**0.70**	**0,054**
Model 2d : 10 (5.6)	**0.76**	**0,023**	**0.67**	**0,060**
Model 2e: 12 (4.6)	**0.76**	**0,023**	**0.65**	**0,058**
Model 2f : 14 (4)	**0.78**	**0,024**	**0.66**	**0,064**
Model 2g : 16 (3.5)	**0.78**	**0,020**	**0.66**	**0,050**
Model 2h : 17 (3.2)	**0.79**	**0,024**	**0.66**	**0,062**
Mean of the 8 means (SD)	**0.75 (0.041)**		**0.67 (0.022)**	

*See Table 1 for abbreviations; # number of variables with, in parentheses, the events/variable ratio

Between-model variability according to number of variables

For the second series of analyses, the results are only presented for stratified samples where the events per variable ratio is fixed across samples (see Table 3). For increasing number of variables the mean AUC tends to increase, as the model is more explicative. However, The variance of the mean AUCs (variance of all the 8 means) in training is slightly higher than the variance due only to sample fluctuations (0.041 versus 0.022), without trend observed with the number of variables. Considering the within model variability of coefficients across the 50 samples (i.e. sample variance), we found that for every predictive variable, the sample variance was systematically lower for the model including 4 variables compared with the model including 17 variables.

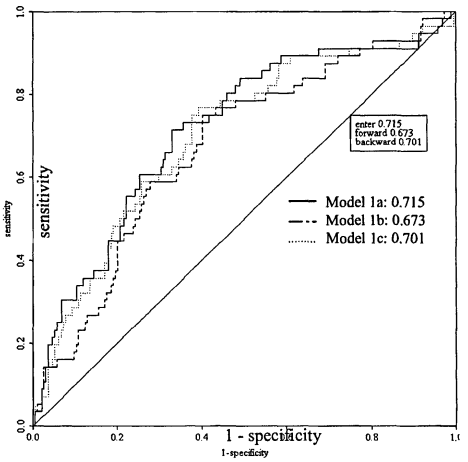

Figure 1: ROC curves for model 1a-1c, in a stratified training sample

Reproducibility of models

The ratio of the mean AUC in train over test samples is 1.081 (standard deviation of 0.11) for model 1a, suggesting a rather reproducible but variable accuracy.

The mean AUC tends to increase in training samples as the number of events per variable decreased, whereas it does not in test samples: the ratio of mean AUC in training samples over AUC in test samples regularly increases from 1.057 to 1.204 while the number of variables increased from 4 (model 2a) to 17 (model 2h), reflecting a decreasing reproducibility while the number of events per variable increases. However, the variability of the discriminative performance is not modified, as reflected by the stable variance of the mean AUCs across the eight models.

Discussion and conclusion

In a series of articles on clinical prediction rules, the Evidence Based Medicine in Critical Care Group, emphasized the importance of evaluating the validity and reliability of these rules for their use in clinical practice [17]. Among questions which should be addressed are "how well does the model categorize patients into different levels of risk" and "how confident are you in the estimate of risk". This implies, for the first one, to measure discrimination and calibration performances of the rule, and for the second one, to measure its precision, and reproducibility. Our main objective was to test for potential causes of instability of the logistic models that could be *a priori* corrected, in order to draw valid conclusions from a validation based on data splitting. Any interpretation of our results should of course be understood restricted to the database which is used for the analyses. In this database of 376 coronary stenoses observed in 84 patients, we considered the stenosis as the statistical unit which is acceptable insofar as only morphological characteristics pertaining to the stenosis are considered as predictive variables. However, we cannot completely eliminate some patient effect in the relationship between these characteristics and clinical characteristics of the patient [18].

This work shows an overall low variability of discriminative performance of logistic models whatever the conditions of modeling are. Within-model standard deviations of AUCs in training samples are never above 0.024 in every set of analyses. However, they are consistently higher (at least twofold) when assessed in test samples, suggesting that variability of predictions can be a problem while a model is used for individual risk prediction in clinical practice. Moreover, the inter-quartile range of AUCs, reflecting random error, corresponds to differences which are often used to conclude to the superiority of a method over another [5]. The principle of a simple data splitting to quantify the predictive performance of a model may therefore be insufficient without some kind of iterative sampling procedures.

In our analyses, decreasing the number of events per variable does not affect the variability of discriminative performance, whereas the analysis of coefficients' variance is in favor of an influence on coefficients' variability. These findings are consistent with the results of Peduzzi and Concato who showed that the number of events per variable should ideally be above 10, or at least above 5, in order to prevent from instability of coefficients in logistic and Cox regression models [14]. The simulation method used by Peduzzi et al. consisted in decreasing the number of events and keeping the same model with the same variables. In our analyses, we kept the same number of events but increased gradually the number of predictive variables, which is somehow closer to the actual modeling practice, seeking for the best model, i.e. the most predictive one without overfitting. This difference in methods prompts us to be careful in drawing clear cut conclusions. Increasing the number of variables influenced the variability of models' regression coefficients, but did not alter the stability of predictive performance. This way of exploring the influence of number of events per variable on stability, brings the problem back to the issue of variable selection, which should be explored further.

We also propose to assess the reproducibility of models using data splitting method and the AUC on test samples. Justice et al. define *reproducibility* as the degree to which the model fits to real patterns in the data and not to random noise, assessed in patients who were not included in learning data but are drawn from the same population [4]. We observed in our data the decrease of reproducibility as the model became more explicative (including more predictive variables). This trend is weak but it is consistent with the already well described overfitting phenomenon: the more variables you include in your model, the more likely you may fit noise to the model, resulting in unstable and poorly reproducible predictions [13].

From this point, perspectives would be, firstly, to check for variability using the same sampling method as Peduzzi et al. for simulations, secondly, to refine interpretation of our results by performing the same analyses on other clinical data. Moreover, further analyses are needed to explore the calibration component of accuracy.

References

[1] Lowensteyn I, Joseph L, Levinton C et al. Can computerized risk profiles help patients improve their coronary Risk? the results of the coronary health assessment study (CHAS). *Prev Med* 1998;27:730-7.

[2] Chatellier G, Menard J. The absolute risk as a guide to influence the treatment decision-making process in mild hypertension. *Hypertens* 1997;15(3):217-9.

[3] Murray LS, Teasdale GM, Murray GD et al. Does prediction of outcome alter patient management? *Lancet* 1993;341:1487-91.

[4] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515-24.

[5] Long WJ, Griffith JL, Selker HP, D'Agostino RB. A comparison of logistic regression to decision-tree induction in a medical domain. *Comput Biomed Res* 1993;26(1):74-97.

[6] Knuiman MW, Vu HT, Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *J Cardiovasc Risk* 1997;4(2):127-34.

[7] Bottaci L, Drew PJ, Hartley JE et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet* 1997;350:469-72.

[8] Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. *Circulation* 1991;83:356-62.

[9] Assmann G, Schulte H, von Eckardstein A. Hypertriglyceridemia and elevated lipoprotein(a) are risk factors for major coronary events in middle-aged men. *Am J Cardiol* 1996;77(14):1179-84.

[10] Cooper GF, Aliferis CF, Ambrosino R et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med* 1997;9:107-38.

[11] Lapuerta P, Azen PS, LaBree L. Use of neural networks in predicting the risk of coronary artery disease. *Comp Biomed Res.* 1995;28:38-52.

[12] Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993;118(3):201-10.

[13] Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15(4):361-87.

[14] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49(12):1373-9.

[15] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837-45.

[16] Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graphical Stat* 1996;5:299-314.

[17] Randolph AG, Guyatt GH, Calvin JE, Doig G, Richardson WS. Understanding articles describing clinical prediction tools. Evidence Based Medicine in Critical Care Group. *Crit Care Med* 1998;26:1603-12.

[18] Ledru F, Theroux P, Lesperance J, et al. Geometric features of coronary artery lesions favoring acute occlusion and myocardial infarction: a quantitative angiographic study *J Am Coll Cardiol* 1999;33(5):1353-61.

Address for correspondence

Isabelle Colombet

colombet@hegp.bhdc.jussieu.fr