*Journal of Statistical Research* *ISSN 0256 - 422 X*
*2017, Vol. 51, No. 2, pp. 131-144*

# A COMPARISON OF INTERNAL VALIDATION METHODS FOR VALIDATING PREDICTIVE MODELS FOR BINARY DATA WITH RARE EVENTS

Momenul Haque Mondol
*Institute of Statistical Research and Training (ISRT)*
*University of Dhaka, Dhaka 1000, Bangladesh*
*Email: mmondol@isrt.ac.bd*

M. Shafiqur Rahman*
*Institute of Statistical Research and Training (ISRT)*
*University of Dhaka, Dhaka 1000, Bangladesh*
*Email: shafiq@isrt.ac.bd*

SUMMARY

In clinical research, prediction models for binary data are frequently developed in logistic regression framework to predict the risk of patient's health status such as death and illness. However, when the outcome is rare, the maximum likelihood (ML) based standard logistic regression has been reported to show poor predictive performance by providing overfitted model. To overcome this, penalized maximum likelihood (PML) based logistic models are being widely used in risk prediction, however, their predictive performance in validation settings is not well-documented. Several validation approaches, namely split-sample, cross-validation, bootstrap validation and its two variants 0.632 and 0.632+, have been widely used to validate the performance of a prediction model, however, it is also unclear which one of these approaches best for estimating accurate predictive performance of a rare-outcome model. This paper focused on evaluating predictive performance of PML based logistic model in such validation settings in comparison with ML based standard model and identifying the effective validation method. An extensive simulation study was performed by creating several scenarios to reflect modeling in dataset with few events. The results revealed that PML based model showed better performance by reducing overfitting to some extent and increasing discriminatory ability over ML based model, irrespective of validation methods under study. Of the validation methods, regular bootstrap and its variants 0.632 and 0.632+, particularly 0.632+, performed well by providing nearly accurate and stable estimate of the true predictive performance. We also illustrated the methods applying them to cardiac data set with few events.

*Keywords and phrases*: Risk prediction, Firth-type penalized regression, Overfitting, Bootstrap-validation

* Corresponding author

# 1 Introduction

Predictive models are increasingly being used in various areas of clinical research such as cardiology, intensive care medicine, and oncology to predict patient's future health status such as death and illness and thereby facilitating patients and providers make shared decision on future course of treatment (Wyatt, 1995; Moons et al., 2009). Given their importance in clinical prediction research, it is essential to assess their predictive performance before using them in practice (Royston et al., 2009; Altman et al., 2009). Two main aspects of model are usually evaluated: (i) calibration - accuracy of prediction and (ii) discrimination - the ability to distinguish between low and high risk patients (Altman and Royston, 2000). A predictive model is generally performed well in terms of both calibration and discrimination in training data (which is used to develop the model) compared to test data (other than training set), even if the later set consists of patients from the same population (Royston et al., 2009; Steyerberg et al., 2001). This behavior is termed as 'optimism'. The problem of 'optimism' associated with predictive model is very common. Hence several approaches have been proposed for an accurate evaluation of predictive performance in a dataset consisting of subjects other than that used to develop the model, rather than a naive evaluation in training sample.

Two common approaches of validation are: internal and external validation, where the former is based on test data from the same population while the later is based on test data from different but relevant population. Of the two approaches, external validation is considered to be more reliable and accurate (Bleeker et al., 2003; Steyerberg et al., 2001). A model with good predictive performance in external validation setting are claimed to provide reasonably accurate predictions for any other patients from a relevant but different population. This concept is generally referred to as 'validity' or 'generalizabilty', and a model with such quality is said to be validated (Justice et al., 1999; Bleeker et al., 2003). However, the test data from external source is hardly available in practice. Alternatively, some internal validation methods are being widely used to validate the prediction models (Steyerberg et al., 2001). Of them, the straight forward approach is split-sample where the whole data are randomly divided into two parts (often 2:1), of which one is used for training the model and the other for testing it. However, this process is reported to be less efficient to overcome optimism, because subjects from both datasets are quite similar as they are from the same underlying population (Toll et al., 2008). Further, there is no guidelines in split-sample process on what proportion of sample should be in training and test data. However, there are other sophisticated approaches that use resampling techniques such as cross-validation and bootstrap validation (Steyerberg et al., 2003, 2001). The cross-validation is an extension of split-sample. For example, in a $k$-fold cross-validation the original sample is partitioned into $k$ subsets of which one is used to test the model and the remaining $k-1$ subsets are used to develop the model (Efron, 1983). The bootstrap method is comparatively efficient, which draws a large number of sample with replacement from the original sample, and the models are trained using bootstrap sample and validated using original sample (Efron and Tibshirani, 1993). The average over large number of repetitions

is an estimate of predictive performance. Two variants of bootstrap resampling such as 0.632 and 0.632+ are also commonly used for internal validation, where 0.632 is an extension of cross-validation and 0.632+ is a further extension of 0.632 (Efron and Tibshirani, 1997). More details on these procedures are discussed later.

Steyerberg et al. (2001) showed a comparison among the above internal validation techniques based on the predictive logistic regression model for binary data. However, developing a predictive model in the standard logistic regression framework and validating its predictive performance is challenging when the outcome is rare or sample size is small or combination of both (Steyerberg et al., 2000). This is because the standard logistic regression produces overfitted model with poor predictive performance (Ambler et al., 2012; Pavlou et al., 2016). The problem of overfitting is very common when the number of event per variable (EPV) in the model is very low, for example, less than 10 (Peduzzi et al., 1996). However, the requirement of minimum EPV (EPV=10) is often difficult to achieve for small or sparse data or even for large data with rare events. Pavlou et al. (2016) and Rahman and Sultana (2017) explored the use of some penalized methods such as ridge (Cessie and van Houwelingen, 1992), lasso (Tibshirani, 1996), Firth-and logF-type penalized methods (Firth, 1993; Greenland and Mansournia, 2015) in risk prediction for binary data with few events. Of them Firth's penalized method is reported to show good performance by removing overfitting to some extent. Although Rahman and Sultana (2017) explored its use in risk prediction, very limited studies have been conducted to assess its predictive performance in validation settings allowing for optimism correction. Further, although several validation methods have been available, it is unclear which one is the most effective for estimating accurate predictive performance of a rare-outcome model. This paper focused on assessing the predictive performance of the Firth-type penalized logistic regression model for rare events in validation settings in comparison with the ML based standard logistic regression model. In addition, this paper compared between the validation approaches to identify the most efficient one for such model using an extensive simulation study.

The paper is organized as follows. Section 2 describes general methodology including both standard and penalized logistic regression models, methods for assessing the predictive ability, and some internal validation approaches. The simulation study is described in Section 3, and an illustration of the methods using real cardiac data is described in Section 4. Sections 5 ends the paper with a general discussion and conclusion.

# 2 Methodology

## 2.1 Maximum Likelihood based Logistic Regression

Let $Y_i$ $(i = 1, 2, \ldots, n)$ be a binary outcome (0/1) for the $i$th subject, which follows Bernoulli distribution with the probability $\pi_i = \Pr(Y_i = 1)$. The logistic regression model can be defined as

$$\text{logit}[\pi_i(\boldsymbol{\beta} \mid \boldsymbol{x}_i)] = \eta_i = \boldsymbol{\beta}'\boldsymbol{x}_i,$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients of order $(k+1)$, and $\boldsymbol{x}_i$ is corresponding predictor vector $(i = 1, \ldots, n)$. The term $\eta_i = \boldsymbol{\beta}'\boldsymbol{x}_i$ is referred to as 'risk score' or 'prognostic index'.

Predictions can be obtained by putting MLEs $\hat{\boldsymbol{\beta}}$ in the following equation

$$\pi(\hat{\boldsymbol{\beta}} \mid \boldsymbol{x}_i) = \left[1 + \exp(-\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)\right]^{-1},$$

where $\hat{\boldsymbol{\beta}}$ is the solution of the score equation $U(\boldsymbol{\beta}) = \partial \log L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = 0$, with the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log[\pi_i(\boldsymbol{\beta})/(1 - \pi_i(\boldsymbol{\beta}))] + n_i \log[1 - \pi_i(\boldsymbol{\beta})] + C$$

for some constant $C$.

## 2.2 Penalized Likelihood based Logistic Regression Model

In order to remove first order bias $(O(n^{-1}))$ due to small-sample in the MLEs of the regression coefficient, Firth (1993) suggested a modified score equation with a penalty term to the above score equation as:

$$U(\beta_r)^{\star} = U(\beta_r) + (1/2)\ \text{trace}[I(\boldsymbol{\beta})^{-1}\{\partial I(\boldsymbol{\beta})/\partial(\beta_r)\}] = 0, \quad (r = 1, \ldots, k)$$

where $I(\boldsymbol{\beta})^{-1}$ is inverse of the information matrix $I(\boldsymbol{\beta}) = -\partial U(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}$. The penalty term used above is known as Jeffreys invariant prior and its influence is asymptotically negligible. The Firth type penalized MLE of $\boldsymbol{\beta}$ is thus

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \{\ell(\boldsymbol{\beta}) + (1/2)\log|I(\boldsymbol{\beta})|\},$$

where $\ell(\boldsymbol{\beta})$ is the log-likelihood and $|I(\boldsymbol{\beta})|$ denotes the determinant of a square matrix $I(\boldsymbol{\beta})$. This approach is known as bias preventive rather than corrective. Because of the penalty term in the score equation, the regression coefficient shrinks towards zero in comparison with MLE, which may help alleviate overfitting.

## 2.3 Evaluating Predictive Performance

The predictive performance of the model is usually evaluated by quantifying (i) the accuracy of prediction (calibration): agreement between the observed and predicted risk and (ii) the ability of the model to distinguish between low-and high-risk patients (discrimination). The calibration can be quantified by estimating calibration slope (CS) and the discrimination by estimating the area under receiver operating characteristic curve (AUC): a graph of sensitivity (true positive rate) against one minus specificity (false positive rate). The calibration slope can be estimated by re-fitting a logistic model with linear predictor or prognostic index (PI: $\hat{\eta}$) derived from the original model as the only covariate (van Houwelingen, 2000):

$$\text{logit}\left[\pi_i \mid \hat{\eta}_i\right] = \beta_0 + \hat{\eta}_i\beta_{PI}.$$

The estimated slope $\hat{\beta}_{PI}$ is the calibration slope for which $\hat{\beta}_{PI} = 1$ suggests perfect calibration, $\hat{\beta}_{PI} < 1$ suggests overfitting, and $\hat{\beta}_{PI} > 1$ suggests under fitting.

The area under ROC curve (AUC) can be estimated using Mann-Whitney U statistic (Obuchowski, 1997), which is equal to those obtained by using trapezoidal rule. Under U-statistic formulation, the AUC can be defined for a pair of subjects $(i, j)$ corresponding to (event vs non-event) as

$$\begin{aligned} AUC &= \Pr\big[\big\{\pi(\hat{\boldsymbol{\beta}} \mid \boldsymbol{x}_i) \mid y_i = 1\big\} > \big\{\pi(\hat{\boldsymbol{\beta}} \mid \boldsymbol{x}_i) \mid y_i = 0\big\}\big] \\ &= \Pr\big[(\hat{\eta}_i \mid y_i = 1) > (\hat{\eta}_j \mid y_j = 0)\big]. \end{aligned}$$

The AUC then can be estimated as

$$\widehat{AUC} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} I(\hat{\eta}_i, \hat{\eta}_j),$$

where $n_1$ and $n_0$ are the number of subjects with and without the event, respectively, and

$$I(\hat{\eta}_i, \hat{\eta}_j) = \begin{cases} 1 & \text{if } \hat{\eta}_i > \hat{\eta}_j \\ 0.5 & \text{if } \hat{\eta}_i = \hat{\eta}_j \\ 0 & \text{if } \hat{\eta}_i < \hat{\eta}_j. \end{cases}$$

The R package "pROC" was used to estimate the AUC and self written R-code was used for estimating calibration slope.

## 2.4 Internal Validation Approaches

The predictive performance of the models developed in both the ML and PML based logistic regression were evaluated in validation settings using five internal validation approaches, namely split sample, cross-validation, regular bootstrap and its two variants 0.632 and 0.632+, with aim to estimate the test performance of the model more accurately than the apparent performance using the training sample. In split sample process, we randomly split the original sample into two equal parts (50% each), of which one was used as training set and the other as test set. The performance in the test set is reported as the estimated performance. In the $k$-fold cross-validation process, we used 10-fold cross validation ($k$=10) where we randomly split the original sample into 10 equal parts, of which one part (10% of the data) was used for testing the model and the rest of the parts (90%) were used for developing the model. The whole process is repeated 10 times by testing the model with the consecutive 10% and developing the model using the remaining 90%. The estimated performance is the average over 10 test performances. More variants of cross-validation can be created by changing the value of $k$, however, the $k$=10, i.e. 10-fold cross-variation is the most recommended fold for validating prediction model (Efron, 1983; Steyerberg et al., 2001). Another important variant of the cross-validation is 'leave-one out' (the 'jackknife')

where one subject is available for testing the model, which is however difficult to implement here because some performance measures such as calibration slope cannot be estimated from the test data with single subject.

The regular bootstrap re-sampling process started with fitting models in bootstrap sample of the same size as the original sample, selected with replacement from the original sample, and evaluated the model performance both in the bootstrap sample and the original sample. The performance in the bootstrap sample is referred to as bootstrap performance and those in the original sample is referred to as test performance. The difference between the bootstrap and test performances is referred to as optimism. The stable optimism is the average over 100 repetitions of the bootstrap re-sampling process. The optimism is then subtracted from the apparent performance (the performance in the original sample) to obtain the optimism corrected estimated performance. Therefore, the estimated performance is {apparent performance − average(bootstrap performance − test performance). Further, two variants of the bootstrap re-sampling method, namely 0.632 and 0.632+, were studied. In the 0.632 process, on an average 63.2% of all subjects in the original sample are at least once selected in the bootstrap sample. The model fitted to the bootstrap sample (consists of 63.2% subjects) was evaluated in the remaining 36.8% subjects. This process is referred as the direct extension of cross-validation where the evaluation of model performance was based on independent test sample. The process is repeated 100 times. Then the estimated performance is the weighted average of the apparent performance (average over 100 estimates in the training sets) and the test performance (average over 100 estimates in the test sets), i.e. the estimated performance is $\{0.368 \times \text{apparent performance} + 0.632 \times \text{test performance}\}$. The other bootstrap re-sampling variant, 0.632+ process, is an extension of 0.632 process with different weighting scheme depending on amount of overfitting (Steyerberg et al., 2001). The estimated performance is defined as $\{(1-w)\times\text{apparent performance}+w\times\text{test performance}\}$, where $w$ can be calculated by relative overfitting $R$ as $w = 0.632/(1 - 0.368 \times R)$, where

$$R = \frac{(\text{test performance} - \text{apparent performance})}{(\text{'no information' performance} - \text{apparent performance})}.$$

The 'no information' performance is the average of the performance in the original sample, where outcome was randomly permuted in each of the 100 replications. For example, the average no information performance for the AUC is 0.5, which is actually same to the performance estimated for the null model (model without covariates). For more details on such approaches, see elsewhere (Steyerberg et al., 2001).

# 3 Simulation Study

A simulation study was conducted to validate the predictive models for binary data with rare outcome using some internal validation approaches. Several simulation scenarios were created by varying the number of events per variable (EPV=3, 5 and 10) in the model to reflect modeling in a dataset consisting of few events. Under each EPV scenario, we first

independently generated a mixture of some continuous and binary covariates, each for a sample of size $n = 300$. The binary covariates $X_1$ and $X_2$ were independently generated from Bernoulli distribution with probability 0.6 and 0.2, respectively, reflecting the scenarios with high or low proportion of subjects being exposed. The continuous covariates $X_3$, $X_4$, $X_5$ were independently generated from the standard normal distribution. Finally, the corresponding binary responses $Y$ were generated from Bernoulli distribution with probability $\pi$ generated from the following model:

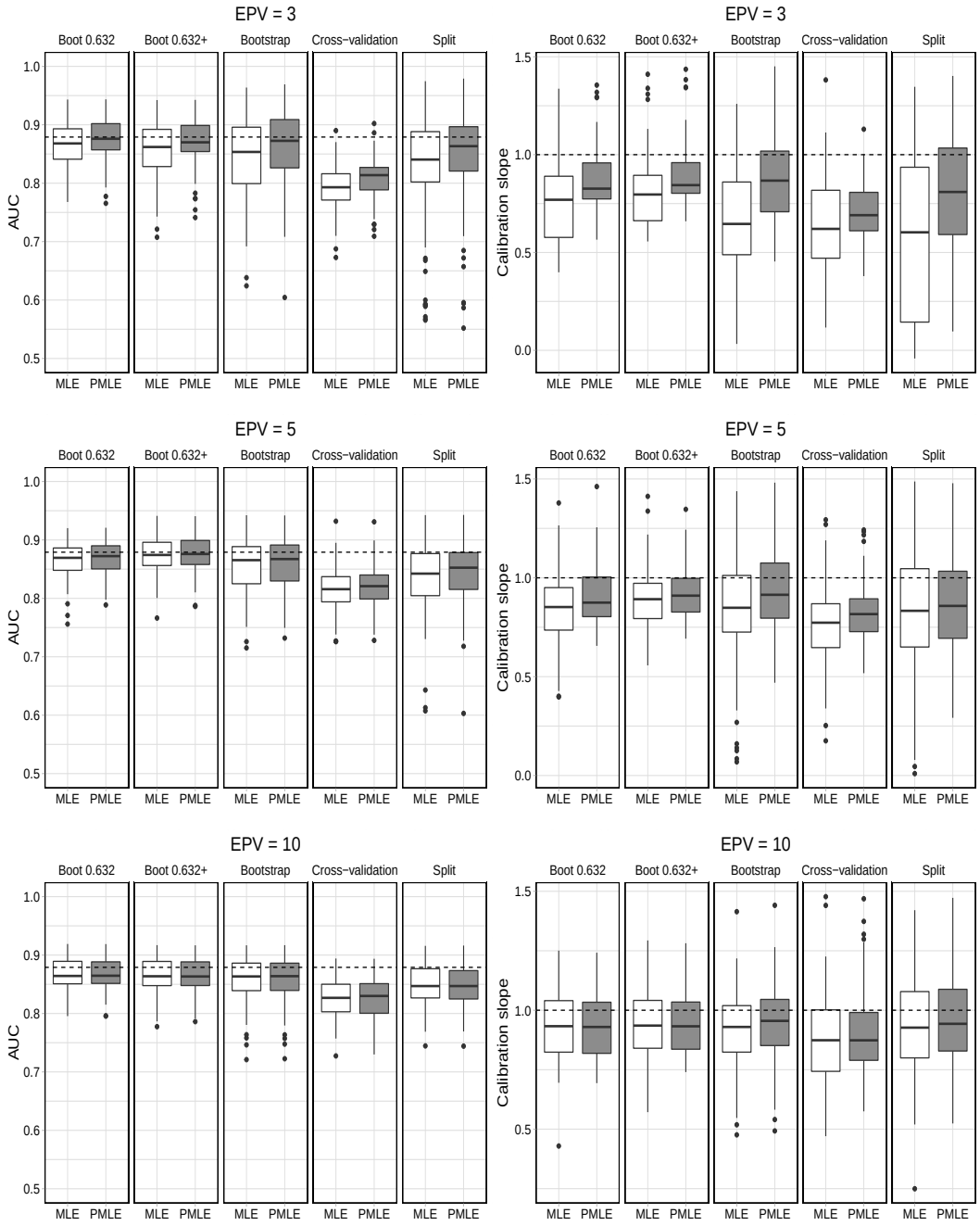
$$\Pr[Y_i = 1 \mid \boldsymbol{x}] = \pi_i = \left[1 + \exp\{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i})\}\right]^{-1}.$$

The true value of the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)'$ were respectively fixed as $(1.005, -0.99, 1.007, 1.01, -1.02)'$. The true value of the intercept $\beta_0$ was varied as -4.6, -3.8, and -2.8 for generating the binary responses with different prevalence so that we have dataset with EPV 3, 5, and 10, respectively. Under each EPV scenario, the above data generating process was repeated to create 100 replications of the data. For each dataset, we evaluated models using the internal validation approaches discussed earlier by creating training and test samples. The logistic regression models based on both MLE and PMLE were fitted in training sample and validated their predictive performance using test sample. In each simulation, the same procedure was repeated as much as required by 10-fold cross-validation (10 repetitions) and bootstrap validation (100 repetitions). The R package "logistf" was used to implement PMLE approach and "glm" function to implement MLE. The R-codes used in this paper can be obtained on request from the first author. The AUC and calibration slope were then used to assess the predictive performance (discrimination and calibration) of the models. For a sensible model the optimal value for calibration slope is, by definition, 1. However, the optimal AUC value was calculated for the true model using a large datasets ($n = 10,000$) with rare outcome (prevalence 5%).

The estimated performance of the respective model for all validation approaches under study were then summarized using box plot. When the models were evaluated using calibration slope, the PMLE based prediction model showed better performance, particularly for low EPV (EPV=3), by improving overfitting to some extent over the MLE based model with the estimated calibration slope more closer to the optimal value 1 indicated by the horizontal dash line (Figure 1). The calibration performance increased with increasing EPV value for both models. These results hold true for all validation approaches. When discriminatory ability of the models were assessed using AUC, PMLE also showed little improvements in discrimination over MLE by providing slightly grater AUC value particularly for low EPV, which is true for all validation approaches (Figure 1).

Of the validation approaches, the bootstrap (regular) and its variants bootstrap 0.632 and bootstrap 0.632+ provided better performance than those for both the cross-validation and split sample. The amount of bias in the estimated performance (difference from the horizontal line indicating optimal performance) induced by all bootstrap re-sampling methods were smaller than those induced by both the split sample and cross-validation approaches. Of them, bootstrap 0.632+ showed the lowest variation in the estimated performance. Both split-sample and cross-validation underestimated the true performance of the model.

Figure 1: The optimism corrected AUC and calibration slope for the models based on MLE and PMLE under different EPV scenarios and validation approaches under study. The horizontal dash line indicates optimal value of the performance measure.



# 4 Illustration using Stress Echocardiography Data

The methods were illustrated using a stress echocardiography data set with rare cardiac events. The dataset is freely available in a public domain 'https://goo.gl/p9VTvL', which is maintained by the Department of Biostatistics, Vanderbilt University, USA. The data were originally extracted from the study conducted by Krivokapich et al. (1999). The aim of the study was to quantify the performance of dobutamine stress echocardiography (DSE) in predicting cardiac events in a total of 558 patients (male 220 and female 338) with known or suspected coronary artery disease. The responses of interest are whether or not a patient developed either of the following events such as 'death due to cardiac arrest', 'myocardial infarction (MI)', 'revascularization by percutaneous transluminal coronary angioplasty (PTCA)' and 'coronary artery bypass grafting surgery (CABG)' over a year following the test. Of a total of 558 patients, 24 patients experienced 'cardiac death', 28 MI, 27 PTCA, 33 CABG, and 89 any cardiac event (any event), indicating rare outcome and low EPV. The main predictor of interest are age, history of hypertension (HsTofHT: yes/no), history of prior MI (HsTofMI: yes/no), status of DSE test (DSE: positive/negative), wall motion anomaly on echocardiogram (restWMA: yes/no), ejection fraction on dobutamine (Dobutamine EF), and ECG (ECG: normal/equivocal). For more details on variables and data description, see elsewhere (Krivokapich et al., 1999; Rahman and Sultana, 2017).

The aim of this illustration is to develop risk prediction models in the logistic regression framework with both MLE and PMLE to predict the risk of having a cardiac event and then to evaluate and compare their predictive performance in internal validation settings. Before performing this, we fitted both MLE and PMLE based logistic models with the predictors of interest mentioned earlier to the original data to examine whether both approaches provide similar estimates of the regression coefficients or not. The selection of predictors was based on the model estimated in the study conducted by Krivokapich et al. (1999) and some exploratory analyses (results not shown). As there were four different cardiac events (MI, PTCA, death, CABG), we fitted separate models for each of them and a model for any of the events denoted by 'anyevent'. From the results in Table 1, it can be observed that the PMLE provided smaller estimates with lower standard error than those with MLE. This is expected as PMLE is a bias preventive approach for small sample case.

Now to develop predictive models with rare cardiac events and assess their predictive performance, we developed predictive model in both MLE and PMLE based logistic regression framework using training sample and assessed their predictive performance in test sample, created for each of the validation settings mentioned earlier. The estimates of the calibration slope and the AUC calculated in test sample of each validation setting were reported as the estimated performance of the models. The results in Table 2 revealed that the bootstrap and its variants, bootstrap 0.632 and bootstrap 0.632+, provided better predictive performance, particularly by improving overfitting, than the other two in all cases. The PMLE based model showed better predictive performance (in terms of both calibration and discrimination) than the MLE based model for the cardiac events with low EPV (<10), irrespective of the validation procedures. However, both the MLE and PMLE showed similar

Table 1: The estimated regression coefficients of the MLE and PMLEs based models. The values in the parenthesis are the standard errors of the estimates.

| | Cardiac events | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MI | | Death | | CABG | | PTCA | | Anyevent | |
| Covariates | MLE | PMLE | MLE | PMLE | MLE | PMLE | MLE | PMLE | MLE | PMLE |
| DSE+ | 0.540 | 0.523 | 0.797 | 0.771 | 1.110 | 1.060 | 0.567 | 0.556 | 0.943 | 0.922 |
| | (0.434) | (0.424) | (0.487) | (0.473) | (0.421) | (0.410) | (0.442) | (0.431) | (0.275) | (0.272) |
| restWMA | -0.823 | -0.764 | -0.522 | -0.502 | -1.086 | -0.994 | -0.419 | -0.379 | -0.749 | -0.734 |
| | (0.632) | (0.604) | (0.583) | (0.564) | (0.687) | (0.647) | (0.597) | (0.576) | (0.354) | (0.349) |
| Dobutamine EF | -0.0352 | -0.0342 | -0.0156 | -0.0157 | -0.0566 | -0.0544 | -0.0120 | -0.0120 | -0.0323 | -0.0315 |
| | (0.0153) | (0.0149) | (0.0173) | (0.0168) | (0.0149) | (0.0145) | (0.0163) | (0.0159) | (0.0105) | (0.0104) |
| Age | 0.0107 | 0.00968 | 0.0308 | 0.0293 | 0.00991 | 0.00890 | 0.00194 | 0.00129 | 0.00551 | 0.00523 |
| | (0.0186) | (0.0183) | (0.0207) | (0.0204) | (0.0180) | (0.0177) | (0.0184) | (0.0180) | (0.0114) | (0.0113) |
| HsTHT | 1.344 | 1.196 | 1.531 | 1.318 | 0.638 | 0.571 | 0.393 | 0.334 | 0.796 | 0.764 |
| | (0.632) | (0.588) | (0.750) | (0.676) | (0.489) | (0.471) | (0.490) | (0.471) | (0.315) | (0.309) |
| HsTMI | 0.396 | 0.388 | 0.0482 | 0.0657 | -0.544 | -0.507 | 1.214 | 1.174 | 0.395 | 0.391 |
| | (0.425) | (0.415) | (0.471) | (0.457) | (0.431) | (0.420) | (0.448) | (0.437) | (0.273) | (0.269) |
| ECG+ | 0.304 | 0.289 | -0.743 | -0.702 | 0.433 | 0.411 | 0.796 | 0.753 | 0.339 | 0.333 |
| | (0.425) | (0.414) | (0.461) | (0.447) | (0.419) | (0.408) | (0.453) | (0.440) | (0.265) | (0.261) |
| Intercept | -2.877 | -2.601 | -5.303 | -4.873 | -0.794 | -0.670 | -3.715 | -3.446 | -1.012 | -0.973 |
| | (1.687) | (1.642) | (1.978) | (1.920) | (1.606) | (1.571) | (1.644) | (1.604) | (1.033) | (1.019) |

performance for cardiac event (anyevent) with high EPV (>10).

# 5 Discussion and Conclusion

Developing and validating a prediction model for rare binary outcome is challenging, because the standard ML based logistic regression showed poor predictive performance by providing overfitted model. This paper explored the use of penalized logistic regression in risk prediction for rare outcome and evaluated its predictive performance in validation settings. In addition, this paper compared some well known internal validation methods, namely split-sample, 10-fold cross-validation, bootstrap validation and its two variants 0.632 and 0.632+ to identify the most efficient one for validating rare-outcome models. The findings of the study revealed that penalized logistic model (PMLE) showed some improvements in both calibration and discrimination when EPV is low, particularly by removing over-fitting to some extent over MLE based standard logistic model, regardless of validation methods. The improvement in calibration is higher than those in discrimination. The reason, as explained by Pavlou et al. (2016) and Rahman and Sultana (2017), is that the PMLE tends to shrink the predicted probability towards the average compared with the MLE and hence the ordering of the predicted probabilities with and without experiencing the event in most patient pairs tend to remain unchanged after shringkage, which resulted in small improvement in

Table 2: The estimated performance of the MLE and PMLE based predictive models quantified using AUC and calibration slope in different internal validation settings

| | | | AUC | | Calibration slope | |
|---|---|---|---|---|---|---|
| Model | EPV | Validation procedures | MLE | PMLE | MLE | PMLE |
| MI | 3 | Split sample | 0.6956 | 0.7341 | 0.5999 | 0.6696 |
| | | Cross-validation | 0.6926 | 0.6939 | 0.3926 | 0.4379 |
| | | Bootstrap (Regular) | 0.7519 | 0.7536 | 0.7535 | 0.8203 |
| | | Bootstrap 0.632 | 0.7504 | 0.7523 | 0.7549 | 0.7730 |
| | | Bootstrap 0.632+ | 0.7356 | 0.7385 | 0.7764 | 0.7893 |
| Sudden death | 3 | Split sample | 0.6900 | 0.6926 | 0.5461 | 0.6055 |
| | | Cross-validation | 0.6119 | 0.6146 | 0.2474 | 0.2492 |
| | | Bootstrap (Regular) | 0.7001 | 0.6984 | 0.6487 | 0.7611 |
| | | Bootstrap 0.632 | 0.6444 | 0.6947 | 0.6810 | 0.6930 |
| | | Bootstrap 0.632+ | 0.6664 | 0.6626 | 0.6871 | 0.7004 |
| CABG | 4 | Split sample | 0.7314 | 0.7349 | 0.6446 | 0.7054 |
| | | Cross-validation | 0.6955 | 0.6968 | 0.5845 | 0.5781 |
| | | Bootstrap (Regular) | 0.8170 | 0.8174 | 0.8213 | 0.8942 |
| | | Bootstrap 0.632 | 0.8026 | 0.8036 | 0.7960 | 0.8214 |
| | | Bootstrap 0.632+ | 0.8096 | 0.8102 | 0.8367 | 0.8547 |
| PTCA | 3 | Split sample | 0.7733 | 0.7793 | 0.8606 | 0.8751 |
| | | Cross-validation | 0.6312 | 0.6301 | 0.2371 | 0.3070 |
| | | Bootstrap (Regular) | 0.7238 | 0.7256 | 0.7582 | 0.7933 |
| | | Bootstrap 0.632 | 0.7236 | 0.7248 | 0.7422 | 0.7507 |
| | | Bootstrap 0.632+ | 0.7157 | 0.7171 | 0.7751 | 0.7821 |
| Any cardiac event | $> 10$ | Split sample | 0.7750 | 0.7754 | 1.0875 | 1.0935 |
| | | Cross-validation | 0.7122 | 0.7126 | 0.8351 | 0.7846 |
| | | Bootstrap (Regular) | 0.7670 | 0.7673 | 0.9018 | 0.9156 |
| | | Bootstrap 0.632 | 0.7670 | 0.7673 | 0.9046 | 0.9047 |
| | | Bootstrap 0.632+ | 0.7713 | 0.7715 | 0.9286 | 0.9279 |

AUC values of the PMLE in comparison with the MLE. The findings are similar to those found in the study conducted by Rahman and Sultana (2017).

When comparing the validation methods, the results of the study revealed that regular bootstrap method and its two variants 0.632 and 0.632+ showed better performance by providing smaller amount of bias in the estimate of the predictive performance (calibration slope and AUC) compared to those associated with the split-sample and cross-validation, for all EPV scenarios. Of them, bootstrap 0.632+ provided the estimate of the predictive performance with lower variability. The split-sample method provided large amount of bias and unstable estimate of the predictive performance for low EPV scenarios. The amount of bias decreased with increasing EPV. On the other hand, although cross-validation has an intuitive understanding as a straight forward extension of simple split-sample method, the 10-fold cross-validation under study provided large amount of bias, even larger than

those associated with split-sample. However, the variability in the estimate is lower, even equal to those associated with the bootstrap re-sampling methods. The probable reason for poor performance of the 10-fold cross-validation is that the test data set with 10% of samples does not have enough events (as we dealt with rare event) to have accurate estimate of the performance measures (calibration slope and AUC). Another reason, as discussed in the literature (Efron and Tibshirani, 1993; Steyerberg et al., 2001), is that this form of cross-validation are not expected to perform better than bootstrap, since the bootstrap was proposed as an improvement over the jack-knife (leave-one out cross-validation), which was not considered here because estimation of the calibration slope and AUC is not possible for data with single subject. These findings are similar to those in the study conducted by Steyerberg et al. (2001), who explored the effectiveness of bootstrap validation methods for the model with frequent events.

The general idea of validating a prediction model is to evaluate its predictive performance using data from external sources but relevant population (Bleeker et al., 2003; Steyerberg et al., 2001). However, the availability of data for external validation is often difficult. For such situation, internal validity is widely considered as an approximation of external validity (or generalizability). However, the evaluation of the model for rare outcome should not be based on typical split-sample based internal validation, as it is reported to underestimate the true predictive performance. Similarly, although the cross-validation has intuitive sense, it may not be recommended for validating model for rare outcome as it produced bias in the estimate of the predictive performance. However, the bootstrap re-sampling method or its variants 0.632 and 0.632+ may be recommended to validate a predictive model with rare outcome as they are reported to provide reasonably accurate and stable estimate of the predictive performance. Of them, bootstrap 0.632+ is the most efficient one and hence are recommended for validating the model with few events. Further, the penalized logistic model might be appropriate choice over standard logistic model for developing prediction model for rare events.

## References

Altman, D. G. and Royston, P. (2000), "What do we mean by validating a prognostic model?" *Statistics in Medicine*, 19, 453–473.

Altman, D. G., Vergouwe, Y., Royston, P., and Moons, K. G. (2009), "Prognosis and prognostic research: validating a prognostic model," *BMJ*, 338, b605.

Ambler, G., Seaman, S., and Omar, R. Z. (2012), "An evaluation of penalised survival methods for developing prognostic models with rare events," *Statistics in Medicine*, 31, 1150–1161.

Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., and Moons, K. (2003), "External validation is necessary in prediction research: A clinical example," *Journal of Clinical Epidemiology*, 56, 826–832.

Cessie, S. L. and van Houwelingen, J. C. (1992), "Ridge Estimators in Logistic Regression," *Journal of the Royal Statistical Society, Series C*, 41, 191–201.

Efron, B. (1983), "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American Statistical Association*, 78, 316–331.

Efron, B. and Tibshirani, R. (1997), "Improvements on cross-validation: the 632+ bootstrap method," *Journal of the American Statistical Association*, 92, 548–560.

Efron, B. and Tibshirani, R. J. (1993), "An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability, Vol. 57," *New York and London: Chapman and Hall/CRC*.

Firth, D. (1993), "Bias reduction of maximum likelihood estimates," *Biometrika*, 80, 27–38.

Greenland, S. and Mansournia, M. A. (2015), "Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions," *Statistics in Medicine*, 34, 3133–3143.

Justice, A. C., Covinsky, K. E., and Berlin, J. A. (1999), "Assessing the generalizability of prognostic information," *Annals of internal medicine*, 130, 515–524.

Krivokapich, J., Child, J., Walter, D. O., and Garfinkel, A. (1999), "Prognostic Value of Dobutamine Stress Echocardiography in Predicting Cardiac Events in Patients With Known or Suspected Coronary Artery Disease," *J Am Coll Cardiol.*, 33, 708–16.

Moons, K. G., Altman, D. G., Vergouwe, Y., and Royston, P. (2009), "Prognosis and prognostic research: application and impact of prognostic models in clinical practice," *BMJ*, 338, b606.

Obuchowski, N. A. (1997), "Nonparametric analysis of clustered ROC curve data," *Biometrics*, 53, 567–578.

Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., and Omar, R. Z. (2016), "Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events." *Statistics in Medicine*, 35, 1159–77.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996), "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, 49, 1373–1379.

Rahman, M. and Sultana, M. (2017), "Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data." *BMC Medical Research Methodology.*, 17:33.

Royston, P., Moons, K. G., Altman, D. G., and Vergouwe, Y. (2009), "Prognosis and prognostic research: developing a prognostic model," *BMJ*, 338, b604.

Steyerberg, E., Eijkemans, M., Harrell, F. E., and Habbema, J. D. (2000), "Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets," *Statistics in Medicine*, 19, 1059–79.

Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., and Moons, K. G. (2003), "Internal and external validation of predictive models: a simulation study of bias and precision in small samples," *Journal of Clinical Epidemiology*, 56, 441–447.

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M., Vergouwe, Y., and Habbema, J. D. F. (2001), "Internal validation of predictive models: efficiency of some procedures for logistic regression analysis," *Journal of Clinical Epidemiology*, 54, 774–781.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Toll, D., Janssen, K., Vergouwe, Y., and Moons, K. (2008), "Validation, updating and impact of clinical prediction rules: a review," *Journal of Clinical Epidemiology*, 61, 1085–1094.

van Houwelingen, H. C. (2000), "Validation, calibration, revision and combination of prognostic survival models," *Statistics in Medicine*, 19, 3401–3415.

Wyatt, J. C. & Altman, D. G. (1995), "Prognostic models: clinically useful or quickly forgotten?" *BMJ*, 311, 539–541.

*Received: November 5, 2017*

*Accepted: January 8, 2018*