Paper 3872-2019

Using SAS® to Validate Prediction Models

Xiaoting Wu, Department of Cardiac Surgery, University of Michigan, Ann Arbor, MI;

Chang He, The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative;

Donald S. Likosky, Department of Cardiac Surgery, University of Michigan, Ann Arbor, MI; The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative

ABSTRACT

Model validation is an important step in establishing a prediction model. A model validation process quantifies how well the model predicts future outcomes. However, there are very few SAS® programming examples showing the validation process. We previously developed a generalized mixed effect model that predicts peri-operative blood transfusion from patients' characteristics. In this paper, we demonstrate a number of SAS techniques that we used to validate such a model. This prediction model was developed using the GLIMMIX Procedure. The validation methods include calibration using SGPLOT, discrimination using the ROC statement in the LOGISTIC Procedure, and sensitivity analysis with a bootstrapping method using the SAS MACRO language.

INTRODUCTION

Prediction models are widely used in fields of health care, clinical practice, economic and society. However, establishing a prediction model is a very complex process. Steyerberg [1] proposed seven steps for developing prediction models, including 1) problem definition and data inspection, 2) predictors coding, 3) model specification, 4) model estimation, 5) model performance, 6) model validation and 7) model presentation. Among these steps, model validation is critical to assess model performance and ensure a model's capability to predict future outcomes [2].

Model validation is generally performed internally or externally [3, 4]. Common measures for model validation include calibration that shows the agreement between the predictive outcomes versus the observed outcomes, discrimination that checks the concordance between predictions and observations, and bootstrap that validates model performance using repeated sampling technique [5].

We used our blood transfusion prediction model as an example to illustrate the model validation process. Evidence shows that unnecessary blood transfusion is independently associated with adverse outcomes after cardiac surgery [6, 7]. This prediction model aimed to provide patients' blood transfusion risk prior to surgery to facilitate clinicians' decision making [8]. For example, given high predictive transfusion risk, blood conservation modalities could be undertaken to reduce a patient's transfusion risk [9].

Our data comprised more than 20,000 coronary artery bypass grafts procedures from multiple hospitals. The transfusion rate was 36.8%. We developed a prediction model based on a patient's preoperative risk factors including demographic factors and medical conditions. During model development, data was randomly split into a model development dataset and a model validation dataset. We used the development dataset for variable selection and functional form assessment, and the validation dataset to assess model performance. Final model parameter estimates were obtained from the large dataset that combined both development and validation datasets.

Our final prediction model is a generalized mixed effect model using the GLIMMIX Procedure that identified 16 preoperative predictors, and accounted for hospitals as random effects [10]. During model validation, we performed model calibration using SGPLOT, discrimination using the ROC option in PROC LOGISTIC and sensitivity analysis using SAS MACRO. These procedures can be applied to internal or external validation.

CALIBRATION

Calibration demonstrates the agreement between observed outcomes and predictions. To perform calibration, the study population is first divided into risk deciles based on predicted probabilities from models. The expected number of outcomes in each decile is calculated by summing the predicted probabilities in each decile, and the observed number of outcomes in each decile is calculated by summing the number of observed outcomes in each decile. Hosmer-Lemeshow test can be used to compare the observed and expected outcomes [11]. A calibration plot can be presented to demonstrate the agreement between the observed and expected. This plot has the expected rates by deciles on the x-axis, and the observed rates by deciles on the y-axis. A good calibration should lie close to a 45-degree line.

First, we used the OUTPUT statement to obtain predictions from our mixed effect model. Option NOBLUP is used to exclude random effects when calculating the predicted probability for each patient. In this model, we have 16 predictors as listed in the MODEL statement. These predictors were chosen by model selection as well as their clinical relationship with blood transfusion. In this model, subject=STS_hospnpi fits the random hospital effect. We used the STORE statement to obtain the model estimate to "parameter_dat" dataset.

```
/*******output prediction from the mixed effect model ****/
proc glimmix data=mix_model;
   class  bsa4c  (ref="LT1.6") albumin_3c (ref=">3.5") female (ref="0")
ef4cat (ref="60%+") crealst4c (ref="LT0.8") race3c (ref="White") status3c
(ref="Elective") vd3 (ref="No") chf_ (ref="No") pvd_ (ref="No")  cvd_
(ref="No") dialysis_ (ref="No") prior_cv(ref="No") STS_hospnpi;

   model rbc = year age bsa4c albumin_3c  hct_ hct_gt36_ hct_gt39_
hct_gt43_ female ef4cat  crealst4c race3c   status3c vd3      chf_ pvd_
cvd_ dialysis_ prior_cv  /link=logit dist=bin  solution ;
   random int/ subject=STS_hospnpi;
   store parameter_dat;
   output out=pre pred(noblup ilink)=p;

run;
```

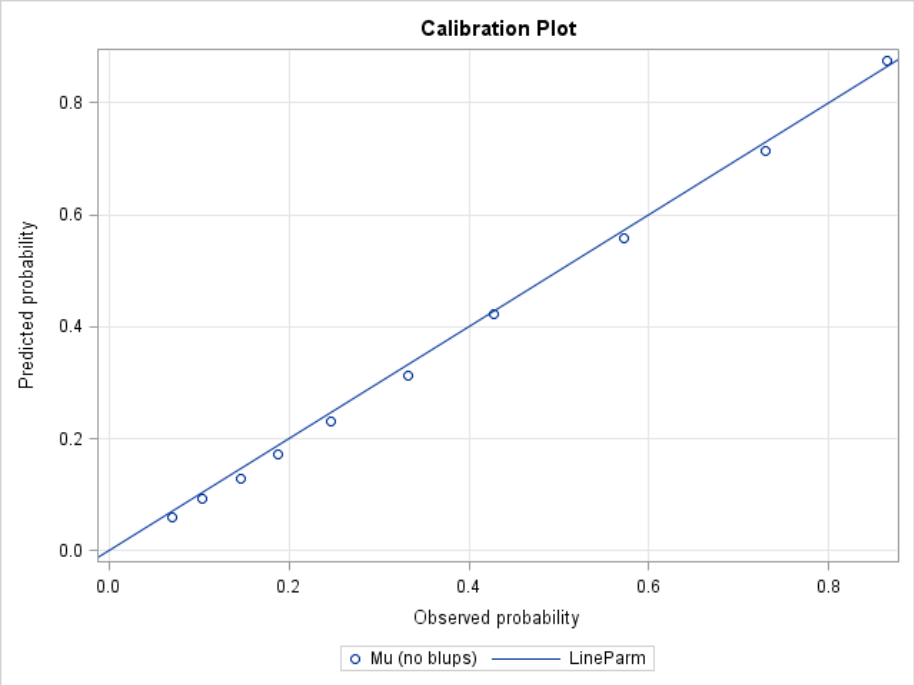
We then ranked the predicted probabilities into deciles using PROC RANK. Variable "p" is the individual transfusion probability calculated by the fixed effect from the model.

```
* create probability deciles, and the ranks of probability <rank_p>;
proc rank data=pre out=ranky descending groups=10;
    var p;
    ranks rank_p;
run;


*output the median probability by deciles ('rank_p');
proc means data=ranky median mean;
   var p ;
   by rank_p;
   output out=median_pr median=median_predict_p mean=mean_predict;
run;

*output the observed transfusion rates by deciles, which is the number of
rbc events divided by the total number of observations in each decile;
proc sql;
   create table observe_pr as
   select sum(rbc) as no_events,  count (*) as no_obs, calculated
no_events/ calculated no_obs as observe_pr, rank_p
   from ranky
   group by rank_p;
quit;

* create the merge dataset that include the median probability and observed
transfusion rate in each decile;

data merge1;
   merge observe_pr median_pr;by rank_p;
run;
```

In each decile, we calculated the observed rate and the median prediction probability. We then plot the calibration plot with the observed transfusion rate in each decile on the x axis, and the median prediction probability on the y axis. The 45-degree line was added as a reference line. The data points which are represented by circles fall close to the reference line (Figure 1). This indicates that the model fits the data well.

```
*Plot the calibration Graph;
proc sgplot data=merge1;
   scatter  x=observe_p y=median_predict_p;
   lineparm x=0 y=0 slope=1; /** plot the reference line **/
   xaxis grid; yaxis grid;
run;
```

Figure 1. Calibration plot



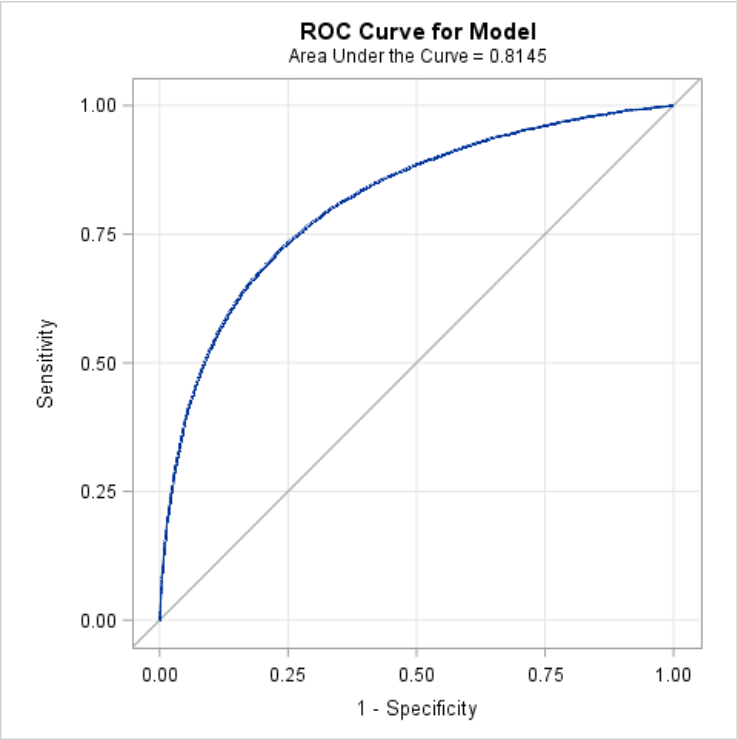
DISCRIMINATION

Discrimination evaluates the ability of a prediction model to discern subjects who had the outcome from subjects who did not have the outcome. A common measure for model discrimination is the area under the receiver operating characteristic (ROC) curve (AUC). The AUC method sets each predicted probability from the prediction model as a threshold and calculates the specificity and sensitivity for each threshold. This is often shown by a ROC curve that plots sensitivity against one minus specificity over all possible thresholds. AUC is equivalent to the c-statistics [12, 13]. C-statistics is a rank order statistic of concordance probability, and is calculated by comparing the predicted probability from randomly selected pairs of subjects with and without the outcome. The C-statistic can be interpreted as the probability that a subject with an observed outcome would have higher probability of predicted outcome than a subject without the observed outcome. A c-statistics between 0.7 and 0.80 usually indicates good models, above 0.8 very good models.

The GLIMMIX procedure does not have a convenient way to directly calculate c-statistics. To obtain the c-statistics from our prediction model, we used the STORE statement in GLIMMIX to store model parameters from our prediction model. Next, we used the PLM Procedure to apply models to a different data set to obtain prediction. We specified the ILINK option on the SCORE statement so that the prediction is at the scale of probability. We applied the predicted probability (variable “Predicted”) to the PROC LOGISTIC model and used the ROC statement to generate the ROC curves (Figure 2).

```
proc plm restore= parameter_dat;
  score data=mix_model out=out/ilink;
run;
proc logistic data=out descending ;
       model rbc = Predicted;
       roc;
       ods output ROCassociation=roc;
    run;
```

Figure 2. ROC curve for model



BOOTSTRAP SAMPLING

To assess model performance in different clinical subgroups, a bootstrap re-sampling method [14] can be used . Bootstrap allows repeatedly estimating a statistic from a large number of bootstrap samples. The bootstrap samples are generated with replacement from the original data [15, 16] . In our example, we generated bootstrap samples by sampling patients with replacement from a defined clinical subgroup. There are many ways to create bootstrap samples in SAS, including the SURVEYSELECT Procedure and do loops. After generating bootstrap samples, we calculated c-statistics in each bootstrapping sample. We then estimated the mean and variance of c-statistics from all bootstrap samples.

As an example, one of the important clinical subgroups for blood transfusion is defined by a patient's admission status (i.e., elective, urgent or emergent). For each admission status, we created 100 bootstrap samples from the original data. We applied the model estimates to these bootstrap samples using PROC PLM RESTORE. We then calculated c-statistics within each sample. From the bootstrap samples, we could obtain standard deviation of the c-statistics. From this sensitivity analysis, we were able to validate how robust our model performance is among different clinical groups.

```
/**create boot samples, part of these codes adapted from Barker et al. ***/
%macro bootsample(b);
data sub1 (where=(status3c="Elective"))
     sub2 (where=(status3c="Urgent"))
     sub3 (where=(status3c="Emergent"); /* Create one data set for each
subgroup */
   set mix_model;
run;

data boot_subgroup;
%do t=1 %to 3;
 do sample=1 to &b;
 do i = 1 to nobs;
 pt = round(ranuni(&t)*nobs) ;  /* ranuni returns a random number from the
uniform distribution on (0,1) interval */
 set sub&t nobs = nobs point=pt;
 output;
 end;
 end;
 %end;
stop;
run;

%mend;

%bootsample(100);
```

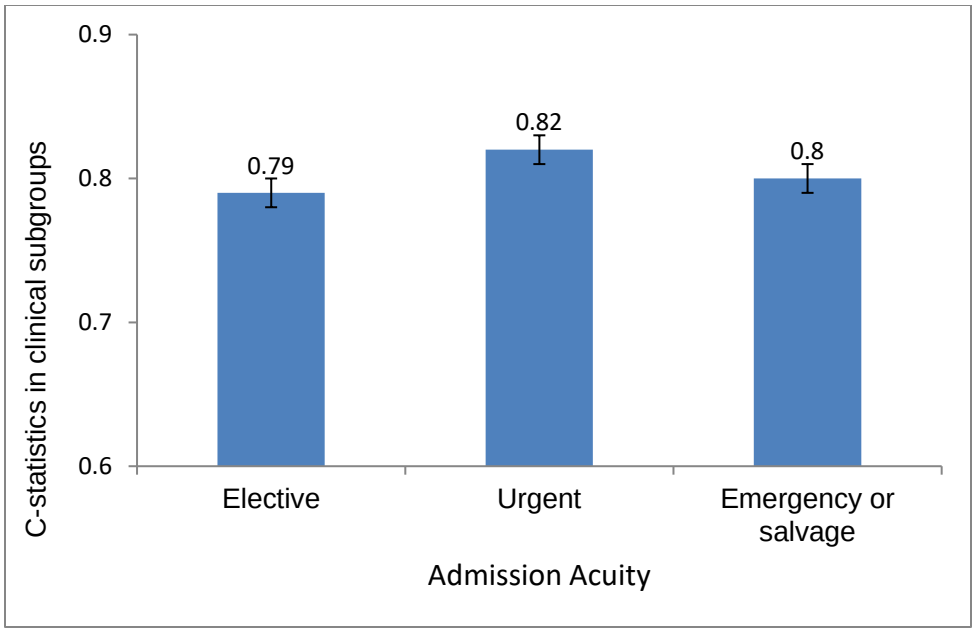
```
/*example: model application to the bootstrapping samples of emergent
status *****/

%macro combine;
%do i=1 %to 100;
  proc plm restore=parameter_dat;
       score data=boot_subgroup(where=(sample=&i and status3c="Emergent"))
out=out&i/ilink;run;

  proc logistic data=out&i  descending ;
        model rbc = Predicted;
        roc;
        ods output ROCassociation=roc&i;
%end;
    run;

 data roc_test;
      set  %do i=1 %to 100;roc&i  %end;
      where ROCModel='Model';
 run;
 %mend;

%combine;


/*** obtain mean and variance for c-statistics of modeling for emergent
status*****/
proc means data=roc_test mean std;
    var area;
run;
```

The model c-statistics was then calculated in different clinical subgroup (Figure 3). For example, here is the result of AUC in emergent admission patients from the bootstrap samples (Table 1).

Table 1. SAS output for bootstrap mean and standard deviation of c-statistics

Analysis Variable : Area Area under the Curve	
Mean	Std Dev
0.8005692	0.0119594

Figure 3. C-statistics in clinical subgroups



CONCLUSION

This paper covers some common techniques for validating the performance of a generalized mixed effect prediction model. We demonstrated SAS applications in model calibration, discrimination and sensitivity analysis (Table 2).

Table 2. Summary of model validation techniques

Model validation techniques	Measure	Interpretation	SAS procedures
Calibration	Calibration Plot	Compares median/mean predicted versus median/mean observed	SGPLOT
Discrimination	c-statistics; ROC curve	Interprets as the probability of correct classification for a pair of subjects with and without the outcome	ROC statement in PROC LOGISTIC
Bootstrap	Bootstrap mean and variance of c-statistics	Sensitivity analysis; estimates the model performance in subgroups	SAS MACRO

REFERENCES

1. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* Eur Heart J, 2014. **35**(29): p. 1925-31.
2. EW, S., *Clinical prediction models: a practical approach to development, validation, and updating.* New York: Springer, 2009.
3. Austin, P.C. and E.W. Steyerberg, *Interpreting the concordance statistic of a logistic regression*

model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol, 2012. **12**: p. 82.
4. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.
5. Frank E.Harrell, J., *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* 2015.
6. Speiss, B.D., *Transfusion and outcome in heart surgery.* Ann Thorac Surg, 2002. **74**(4): p. 986-7.
7. Paone, G., et al., *Transfusion of 1 and 2 units of red blood cells is associated with increased morbidity and mortality.* Ann Thorac Surg, 2014. **97**(1): p. 87-93; discussion 93-4.
8. Alghamdi, A.A., et al., *Development and validation of Transfusion Risk Understanding Scoring Tool (TRUST) to stratify cardiac surgery patients according to their blood transfusion needs.* Transfusion, 2006. **46**(7): p. 1120-9.
9. Ranucci, M., et al., *Predicting transfusions in cardiac surgery: the easier, the better: the Transfusion Risk and Clinical Knowledge score.* Vox Sang, 2009. **96**(4): p. 324-32.
10. Likosky, D.S., et al., *Prediction of Transfusions After Isolated Coronary Artery Bypass Grafting Surgical Procedures.* Ann Thorac Surg, 2017. **103**(3): p. 764-772.
11. Crowson, C.S., E.J. Atkinson, and T.M. Therneau, *Assessing calibration of prognostic risk scores.* Stat Methods Med Res, 2016. **25**(4): p. 1692-706.
12. Duchnowski, M., *Predictive Models: Storing, Scoring and Evaluating* SAS, 2017.
13. James A.Hanley, P.B.J.M., MD, PHD; , *The meaning and use of the area under a receiver operating characterisitic (ROC) curve.* Radiology 1982. **143**: p. 29-36.
14. Efron, B., *1977 Rietz Lecture - Bootstrap Methods - Another Look at the Jackknife.* Annals of Statistics, 1979. **7**(1): p. 1-26.
15. MR, C., *Bootstrap Methods: A Practioner's Guide.* John Wiley & Sons, 1999.
16. Nancy Barker, O.P.S., Wallingford, UK, *A Practical Introduction to the Bootstrap Using the SAS System.* Semantic Scholar, 2005.

ACKNOWLEDGMENTS

Thanks to the support from The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Xiaoting Wu (Ting), PhD, MS

Department of Cardiac Surgery

1500 E Medical Center Drive

Ann Arbor, MI 48109

734.936.7731

xiaotinw@med.umich.edu