PaperAS04

Evaluating Predictive Accuracy of Survival Models with PROC PHREG

DEBRUS Roxane, EMCD, Terumo N.V., Leuven, Belgium
JACOBS Lotte, EMCD, Terumo N.V., Leuven, Belgium

ABSTRACT

Predictive accuracy measures the ability of a model to predict future risks. It can be used to validate a model or to compare new data to a pre-existing model. The concordance statistic (C-statistic) is the most commonly used discrimination measure in the context of logistic regression with binary outcomes. Besides the C-statistic, receiver operator characteristic (ROC) curves and AUC (area under the ROC curve) statistics are also commonly used to assess the discrimination ability of the model with binary outcomes. In the context of survival analysis, various C-statistics have been formulated to deal with right-censored data. In SAS, the PHREG procedure provides a number of state-of-the-art techniques to calculate overall concordance statistics and time-dependent ROC curves and AUC statistics for right-censored data. This presentation describes how to use these criteria to validate and compare fitted survival models and presents an example to illustrate its application.

INTRODUCTION

Terumo Corporation was founded in 1921 in Japan. Since then, Terumo developed more than 100 different medical devices in multiple fields. This paper focus on some of the clinical studies investigating devices in the Terumo Interventional Systems product portfolio (Figure 1). One of the fields is radial intervention, which is an interventional therapy performed to access and visualize lesions inside the arteries of a patient. By accessing the arteries starting from a puncture in the wrist, the procedure is minimally invasive and increases the patient's comfort during and after the procedure. Once the arteries are accessed, a guide wire can be placed to reach the diseased artery that needs to be treated. If the disease is located in the arteries of the heart, we talk about coronary artery disease, if the disease is located in the arteries of any other part of the body, for example arteries of the neck, we talk about peripheral vascular disease. Both disease types can be treated by placing a stent at the location of the lesion to open the artery. One of coronary artery stents in the Terumo product portfolio is the Ultimaster™ stent, which has been thoroughly investigated in many clinical trials.

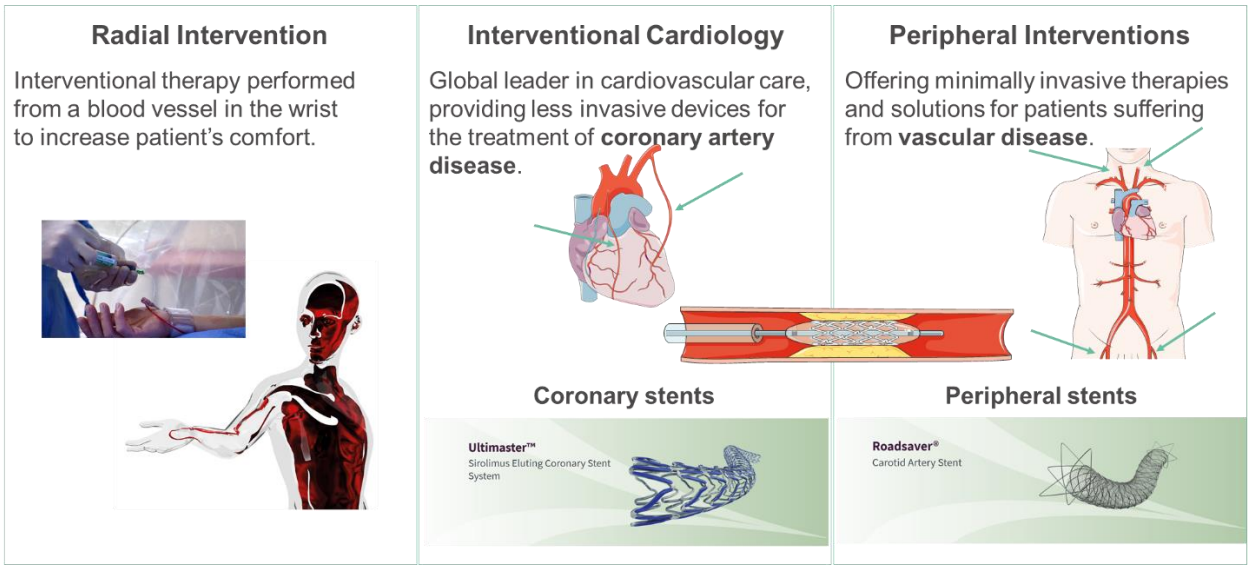

Figure 1 – *Terumo's clinical research fields and product portfolio*

In 2019, Terumo closed its biggest study ever with the Ultimaster™ stent, the e-ULTIMASTER study, which is one of the largest prospective worldwide registries in its field, enrolling up to 37 000 patients from 378 sites across 50 countries. Patients included in the study were treated with a coronary stent because they had a narrowing of the arteries supplying blood and oxygen to the heart. By placing one or more stents at the location of one or more lesion the blood flow to the heart was restored and symptoms were relieved. The patient was followed up for 1 year. The primary endpoint was to validate efficacy and safety of the device based on a composite endpoint of different serious adverse event up to 1 year after the procedure.

In order to evaluate the accuracy of a test that predicts dichotomous outcomes based on a logistic regression model, four main indices are used (Figure 2):

Predicted Risk	Outcome Yes	Outcome No	Total
Positive	50	20	70
Negative	10	120	130
Total	60	140	200

Figure 2 – *Two-ways table comparing predicted risks and outcomes.*

- **Sensitivity**
  Proportion of positive predictive results among people with observed outcome: Se = 50/60 = 0.83
  In other words, *of all the subjects that had an event, what percentage could we have identified as at risk*?
- **Specificity**
  Proportion of negative predictive results in people without observed outcome: Sp = 120/140 = 0.86
  In other words, *of all the subjects that did not have an event, what percentage could we have identified as not being at risk*?
- **Positive Predictive Value**
  Proportion of people having the outcome among those with positive predictive results: PPV=50/70= 0.71
  In other words, *of all the subjects that were identified as being at risk, what percentage did have an event?*
- **Negative Predictive Value**
  Proportion of people without outcome among those with negative predictive results: NPV=120/130= 0.92
  In other words, *of all the subjects that were identified as not being at risk, what percentage did not have an event?*

Based on these, two additional important aspects of a prediction model can be assessed:

- **Calibration** refers to the ability of the model to correctly rank the individuals in the sample by risk.
- **Discrimination** characterizes the model's ability to correctly classify subjects for their actual outcomes. The receiver operator characteristic (ROC) curve, the AUC (area under the ROC curve) and the concordance statistic (or C-statistic) are the most commonly used discrimination measure in the context of logistic regression with binary outcomes (Figures 3-4).

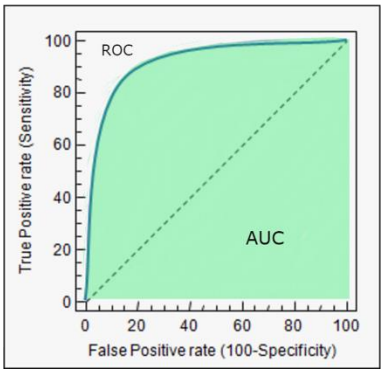

Figure 3 – *ROC curve and AUC*

The higher the C-statistic, the better the model can discriminate between subjects who experience the outcome of interest and subjects who do not.

Figure 4 – *Interpretation scale of the C-Statistic*

In the context of survival analysis modeling, various C-statistics have been formulated to deal with right-censored data. One of these is the ***Harrell's C-index*** also known as the ***concordance index***. It measures the goodness of fit for models which produce risk scores. Whereas C-statistics provide overall measures of predictive accuracy, time-dependent ROC curves and AUC functions summarize the predictive accuracy at specific times.

For a given patient, we are interested in predicting the "time-to-event", which is the time duration until event occurrence, based on some exposure defined as covariates and available at baseline.

BUSINESS CASE

To show an example of how to practically apply these predictive accuracy measures to externally validate an existing risk model in the field of Cardiovascular medicine, let's apply it to a real business case.

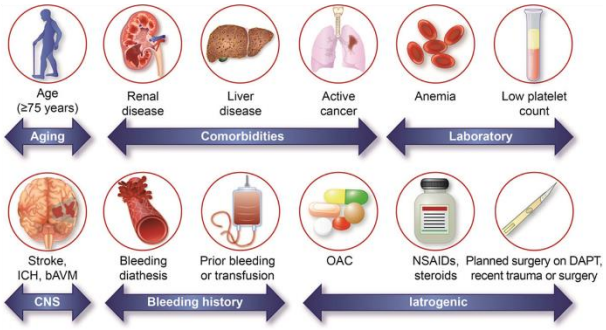

Figure 5 – *Factors associated with an increased bleeding risk after percutaneous coronary intervention*

The existing predictive risk model that we use as an example was developed by an Academic Research Consortium of experts. It was a model specifically designed in a specific population of patients with coronary artery disease who had to be treated with the implantation of a coronary stent. Moreover, these patients were also at high bleeding risk.

Figure 5 illustrates the parameters that classify a patient as high risk of developing a bleeding event after a PCI.

In the patient population it is important to manage the bleeding risk as well as the thrombotic risk. That is the reason why in this specific patient population two risk models were developed, the first one to predict thrombotic events, including myocardial infarction or stent thrombosis, and the second one to predict the risk of major bleedings. The C-statistics were 0.68 and 0.69, respectively (see Figure 6).

Predictor	Thrombotic Risk	Bleeding Risk
Age		■
Diabetes	■	
Prior myocardial infarction		
Liver disease, cancer or surgery		■
Chronic obstructive pulmonary disease		■
Current smoker	■	■
Clinical presentation of disease	■	
Hemoglobin	■	■
Renal function (eGFR)	■	■
Complex procedure	■	■
Bare metal stent	■	
Oral anticoagulation medication		■
C-statistic	**0.69**	**0.68**

Figure 6 – *Multivariate predictors of events identified by both models at 365 days*

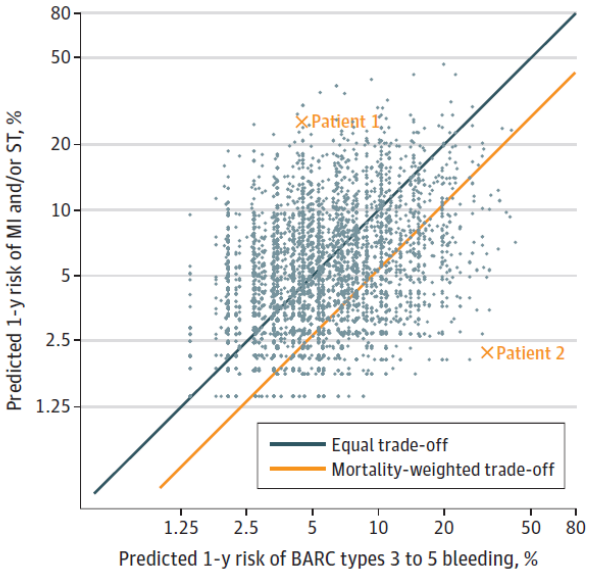

Figure 7 – *Predicted 1-y risk in patients at high bleeding risk*

Finding the right balance between these risks is the ultimate goal of these models, in order to inform and support clinical practice guidelines as to what is the best antithrombotic treatment option in these high bleeding risk patients (Figure 7).

In order to assess the calibration of both models, comparison between observed and predicted risks was performed for patients classified in 5 ordered quintiles of risk, and it showed a good model fit. For both events types, the top quintile had more than 5 times the risk of those in the bottom quintile (Figure 8).

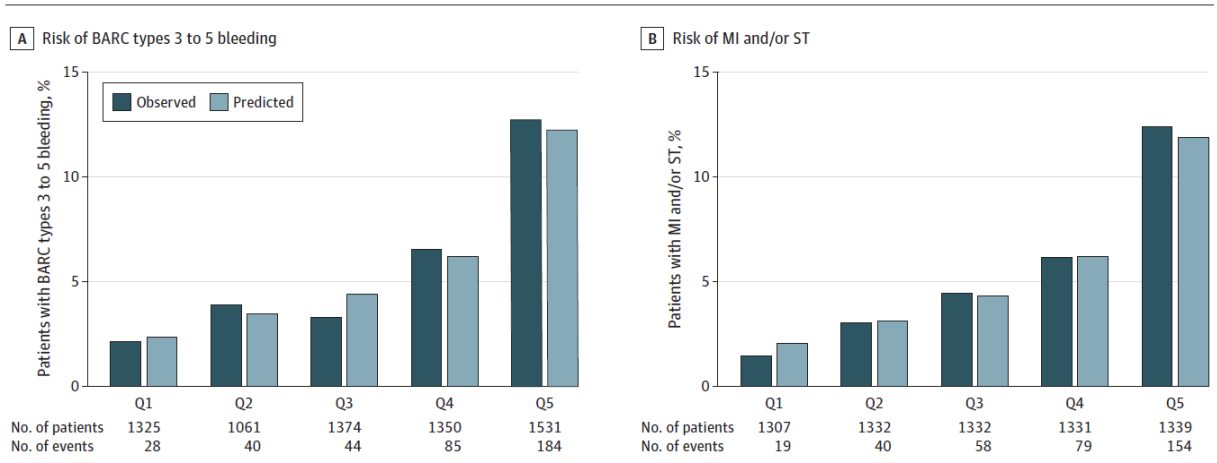

Figure 8 – *Comparison between observed and predicted risks distribution, classified according to quintiles for bleeding events (on the left) and thrombotic events (on the right).*

By reproducing this analysis on Terumo's data, the aim was to validate both bleeding and thrombotic risk models with data from an external study. The large e-UM registry has been used, which included 37198 patients with coronary artery disease that were treated with the Ultimaster coronary stent. Of this population, patients with a high bleeding risk (HBR) were identified in order to have a similar patient population as the one the models were developed from. In the Terumo's registry, not all parameters were available, and 5017 patients with HBR were identified, based on the available ARC-HBR characteristics (i.e., age, renal disease, active cancer, previous stroke, and OAC intake).

The first step was to use the algorithm developed by the ARC-HBR consortium based on baseline hazard functions and hazard ratio to calculate both thrombotic and bleeding Predicted Risks (PR) for each subject (Figure 9).

When a parameter was missing or not collected, it was considered as a "No" if it was a risk factor and as "Normal" if it was an assessment.

By looking at the distribution of both predicted risks (Figure 10), it appears that most of the subjects have a relatively low risk and mean and median are similar, but the maximum value of predicted thrombotic risk (on the left) is much higher than the maximum value of the predicted bleeding risk (on the right).

Thrombotic PR	Response	Assigned value
Diabetes	No	1
	Yes	1.56
Smoker	No	1
	Yes	1.48
Prior MI	No	1
	Yes	1.89
NSTEMI – STEMI Presentation	No	1
	Yes	1.82
Hb	≥130	1
	≥110 and <130	1.27
	<110	1.50
EGFR	≥60	1
	≥30 and <59	1.30
	<30	1.69
Complex	No	1
	Yes	1.50
Bare Metal Stent	No	1
	Yes	1.53

Bleeding PR	Response	Assigned Value
Age	<65	1
	≥65	1.50
Smoker	No	1
	Yes	1.47
COPD	No	1
	Yes	1.39
Hb	≥130	1
	≥110 and <130	1.69
	<110	3.99
EGFR	≥60	1
	≥30 and <59	0.99
	<30	1.43
Complex	No	1
	Yes	1.32
OAC	No	1
	Yes	2.0
Liver cancer surgery	(liver disease=Yes) OR (cancer=Yes and years≤3) OR (surgery=Yes)	1.63
	Otherwise	1

Figure 9 – *Predicted Risk (PR) calculation factors for each subject.*

Predicted thrombotic risk (MI/ST)
(Min= 1.40; Max= 24.7; Mean= 3.72, Median= 3.25)

Predicted bleeding risk (BARC3-5 Bleeding)
(Min= 1.39; Max= 12.52; Mean= 3.63, Median= 3.62)

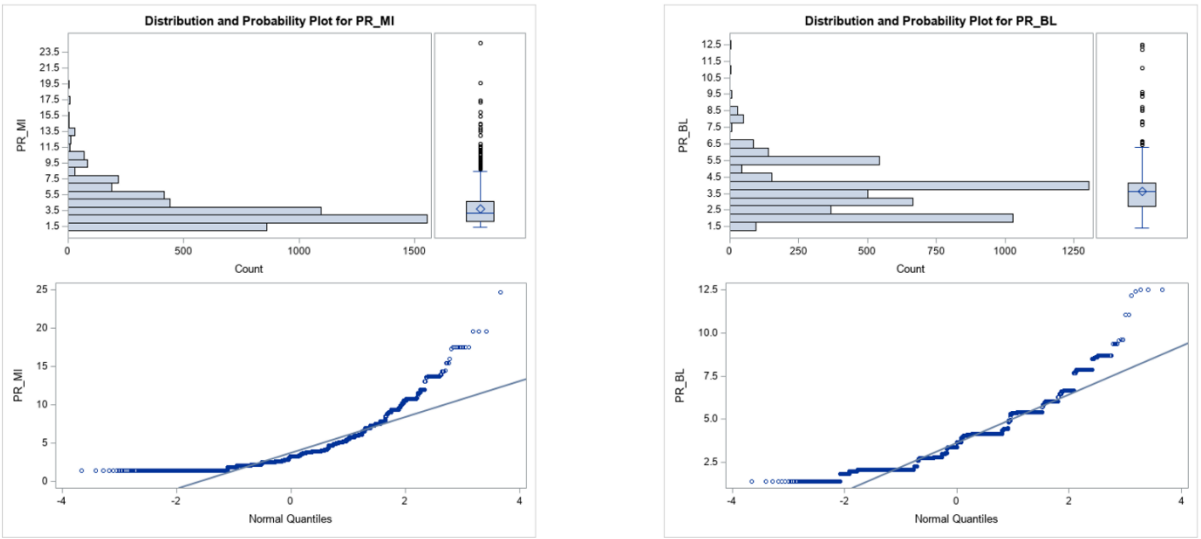

Figure 10 – *Distribution of both Predicted Risk (PR)*

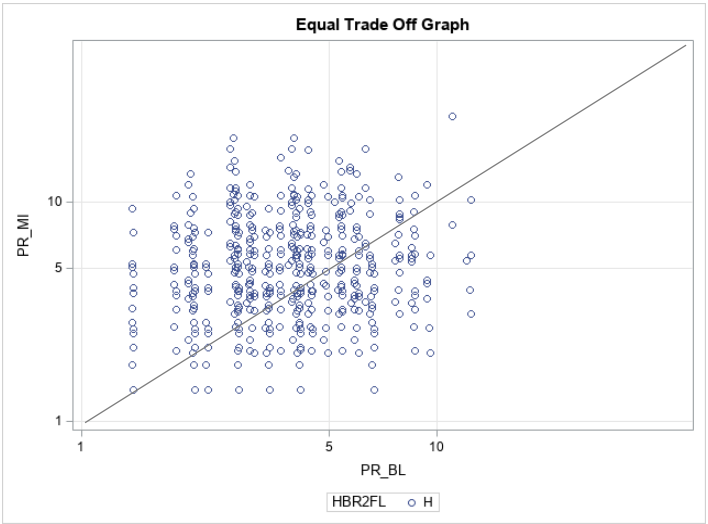

Figure 11 – *Predicted thrombotic risk (PR_MI) versus predicted bleeding risk (PR_BL) in high bleeding risk patients (H).*

When comparing both risks per subject, by setting the predicted risk of bleeding on the x axis, and the predicted risk of thrombotic event on the y axis in a scatter plot using a logarithmic scale, we can observe that these are not equally distributed between the subjects (see Figure 11).

The subjects presented in the upper left part have an increased risk of thrombotic event while the ones in the lower right part have an increased risk of bleeding.

In order to assess the calibration of the model, the full distribution of the predicted risk was classified into quintiles and its mean value was used to assign the reference value for the predicted score. Then for each quintile, the number of events observed was divided by the numbers of patients in order to calculate the observed occurrence of event. This was performed for both models predicting thrombotic event on the left and bleeding event on the right (see Figure 12). There appears to be quite a good fit between predicted and observed risks for the thrombotic event predictive model in most of the classes, the observed occurrence is in the range of the quintile. However, for the bleeding event predictive model, the calibration is less good with marked differences between predicted and observed risk distributions.

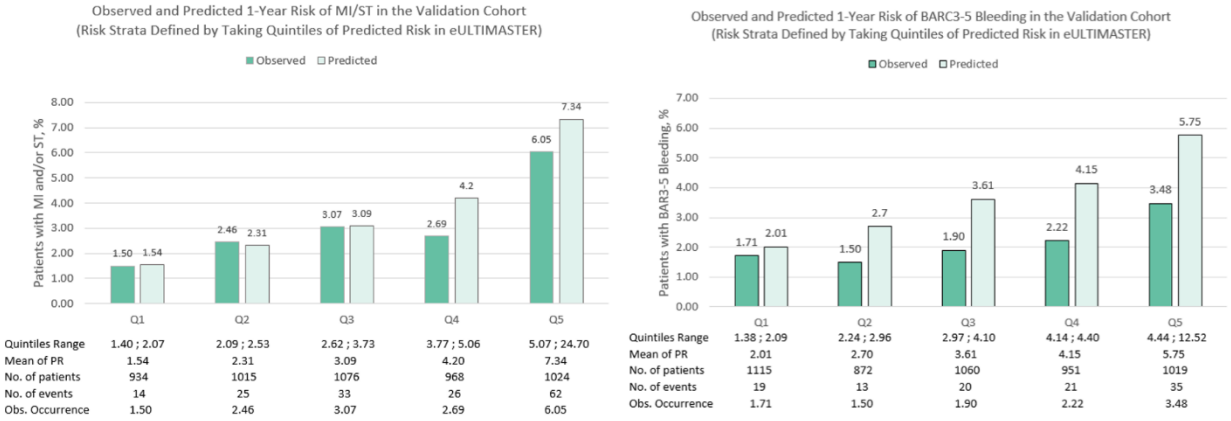

Figure 12 – *Comparison between Observed and predicted risks distribution, classified according to quintiles for bleeding events (on the right) and thrombotic events (on the left).*

The second parameter to assess the performance of the model is the discrimination of the model and in order to do so, the option “Concordance” can be added in the proc phreg statement in SAS in order to calculate the Harrell’s c-statistic while the “plots” option can be used with the “rocoptions” adjustment to define the major timepoints to be used for the analysis. Here there was a particular interest at 1 month, 3 months and 1 year after the procedure. Both results then appear in the output window. As mentioned previously the C-statistic is more global while the ROC curves are time-de pendent and give different result over time (Figure 13).

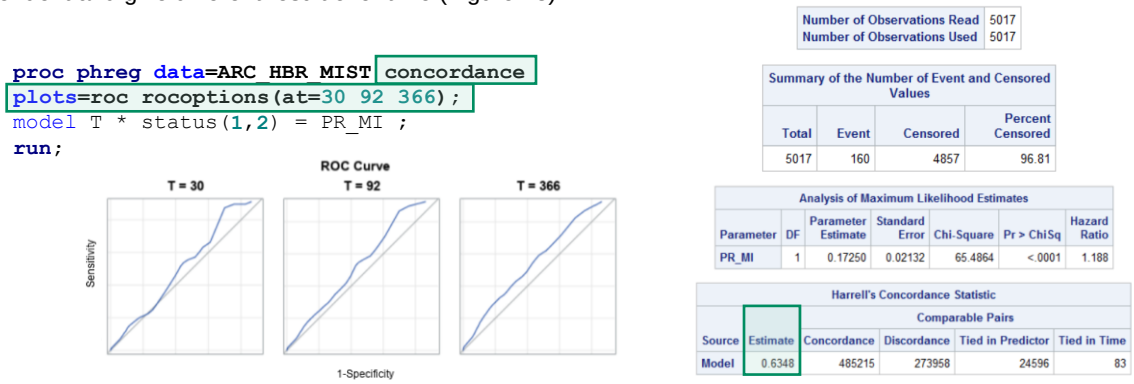

Figure 13 – *SAS Code and output window*

This methodology has been applied to generate two Cox proportional hazards models and evaluate the predictive accuracy of the survival models by using Harrell’s concordance statistic for the high bleeding risk subjects enrolled in the e-Ultimaster study and a C-statistic of 0.64 was obtained for the thrombotic event risk and 0.59 for the bleeding event risk (see Figure 14).

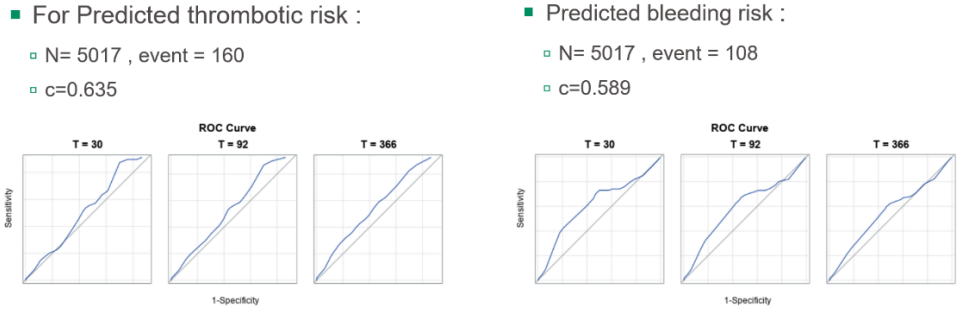

Figure 14 – *Predictive accuracy of both survival models for bleeding events (on the right) and thrombotic events (on the left).*

In case there would be a desire to go one step further, additional programming options are available:

- **Uno's C-Statistic instead of Harrell's C-Statistic**
  - Harrell's C-statistic (1986) : straightforward approach by discarding the pairs that have become non comparable because of censoring → By default CONCORDANCE = HARRELL
  - Uno's C-statistic (2011) : new formulation that models the censoring distribution and uses it to weigh the uncensored observations in the estimation, making the estimates censoring-independent
    → Need to specify : CONCORDANCE = UNO
- **Additional Standard Error (SE) calculation**
  - Based on a perturbation-resampling method proposed by Uno et al. (2011).
  - Need to specify CONCORDANCE = HARRELL(SE) or CONCORDANCE = UNO(SE)
  - Additional options can control this perturbation process : ITER= , SEED=, DIFF
- **ROC Plot Options**
  - PHREG procedure use PLOTS=ROC options to produce ROC curves.
    Additional ROCOPTIONS can be programmed : AT= , AT=*x* TO *y* BY *z*, AUC, AUCDIFF, IAUC, etc.

Even though the 2 models might appear as having a poor accuracy, it is consistent with the score that the main authors obtained, also taking into consideration the limitations of this analysis. In fact, in this study not all the parameters needed to define the patients having a high bleeding risk as per the latest guidelines were collected and therefore a mis-selection could have occurred. On top of that, not all the parameters used by the main authors to generate their risk score were available in the e-Ultimaster database and the lower score was imputed for all missing information.

Finally, it is important to emphasize that the rate of bleeding reported up to 1 year after the procedure is lower than expected, and this might be due to the non-adjudication of this specific event and a possible under-reporting of events.

CONCLUSION

The conventional way to use the C-statistic is when producing a logistic regression model with a binary outcome. The proc phreg procedure can also be used to calculate equivalent predictive accuracy measures (ROC and C-Stat) to validate survival models, and additional options can be used to further finetune the analysis.

The business case presented here shows an example of its use in the field of cardiovascular diseases. Despite the limitations of an external validation cohort, this analysis shows a practical application and the feasibility of externally validate predictive risk models.

REFERENCES

- eUltimaster Terumo Page : https://www.terumo-europe.com/en-emea/clinicaldata/e-ultimaster-trial
- Urban P et al. Defining high bleeding risk in patients undergoing percutaneous coronary intervention: a consensus document from the Academic Research Consortium for High Bleeding Risk. *Eur Heart Journal* 2019;40:2632–2653
- Urban P et al. Assessing the Risks of Bleeding vs Thrombotic Events in Patients at High Bleeding Risk After Coronary Stent Implantation: The ARC–High Bleeding Risk Trade-off Model. *JAMA Cardiol* 2021;**6**:410–419
- Harrell FE. The PHGLM Procedure. In SUGI Supplemental Library Guide 1986, Version 5 Edition, 1986, SAS Institute Inc., Cary, NC, USA
- Uno H et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;**30**:1105-17
- Guo C et al. Evaluating Predictive Accuracy of Survival Models with PROC PHREG. Paper SAS462-2017. SAS Institute Inc., Cary, NC, USA
- SAS/STAT® 14.3 User's Guide. The PHREG Procedure. 2017, SAS Institute Inc., Cary, NC, USA

ACKNOWLEDGMENTS

We wish to thank all former and current colleagues who generously provided their support, insight and perspective that greatly helped to develop the presentation and improve the quality of this paper, in particular: Christophe Giot, Vladimir Borovicanin, Dragica Paunovic, Kanae Hashimoto, Laurence Seidel, Prof. Adelin Albert and all the eULTIMASTER study team.

CONTACT INFORMATION

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

DEBRUS Roxane
Biostatistics Manager EMEA
TERUMO EUROPE NV.
Researchpark Haasrode 1520
Interleuvenlaan 40
3001 Leuven, Belgium
Phone: +32 475 95 35 00
Email: roxane.debrus@terumo-europe.com

JACOBS Lotte
Senior Medical Affairs Specialist
TERUMO EUROPE NV.
Researchpark Haasrode 1520
Interleuvenlaan 40
3001 Leuven, Belgium
Phone: +32 492 25 10 66
Email: lotte.jacobs@terumo-europe.com

Web: https://www.terumo-europe.com/en-EMEA