EuroIntervention

A guide to interpreting and assessing the performance of prediction models

Vasim Farooq[1], MBChB, MRCP; Salvatore Brugaletta[1,3], MD; Pascal Vranckx[2], MD; Patrick W. Serruys[1]*, MD, PhD

1. Thoraxcenter, Erasmus University Medical Center, Rotterdam, The Netherlands; 2. Cardialysis, Rotterdam, The Netherlands; 3.Thorax Institute, Department of Cardiology, Hospital Clinic, Barcelona, Spain

The authors have no conflict of interest to declare.

Risk stratification is an integral and increasingly important aspect of the assessment of patients who are candidates for coronary revascularisation. Careful risk assessment for each patient, based on both clinical and angiographic characteristics, informs decisions regarding aggressive therapeutic interventions, triage among alternative hospital care levels and allocation of clinical resources.

Capodanno et al recently raised the interest within the interventional community on the importance of the assessment of performance of a prognostic score or prediction models.[1,2] The performance of a risk model or prognostic score had however been well established within statistical literature, with up to four different assessments being previously described (Table 1). An understanding of the basic concepts of the assessment of the prediction models are therefore essential, especially since this is currently subject to an intense area of research and new methods to refine these traditional concepts have and are still being developed.[3]

Steyerberg et al[3] recently eloquently summarised these concepts. Traditional measures for binary and survival outcomes include the Brier score to indicate overall model performance, the concordance (or c) statistic for discriminative ability (or area under the receiver operating characteristic ROC curve), and goodness-of-fit statistics for calibration. Consequently, it has been suggested and recommended that, as a minimum, the reporting of discrimination and calibration are essential for understanding the importance of a prediction model, with a recommendation against relying on the c-statistic alone.[3,4]

Table 1. Assessment of the performance of a prognostic score.

1) How accurate is the score as a whole?
i.e., overall performance score
2) How well can the score discriminate between those who do and do not experience the event?
i.e., discrimination measure (e.g., C-statistics)
3) Is the score correctly calibrated?
i.e., goodness-of-fit (e.g., Hosmer-Lemeshow)
4) Is the score transportable or generalisable?

The overall performance of the score

The scale of agreement (or lack of) between the predicted and actual outcomes (i.e., "goodness-of-fit" of the model) are central in allowing the assessment of the overall model performance. The overall model performance essentially captures both calibration and discrimination aspects as discussed below. The distances between observed and predicted outcomes are related to this concept, with better models having smaller distances between predicted and observed outcomes.

One such measure used widely to assess these concepts is the Brier score; this being initially proposed in the 1950s by Glenn Brier as a means to verify weather forecasts in terms of probability.[9] The Brier score is a quadratic scoring rule based on the average squared deviation between predicted probabilities for a set of events and their observed outcomes (Figure 1). Consequently, the score consists of only positive values ranging from 0 (perfect prediction) to 1 (worst possible prediction), with lower scores representing a higher accuracy and no rule-of-thumb, *per se*, on what constitutes an acceptable value. This would potentially allow comparison across different prediction models. [3,5-8]

Discrimination and calibration

Accurate predictions discriminate between those with without the outcome. Individuals are categorised into different outcome groups on the basis of their risk model score in order to allow the physician

** Corresponding author: Thoraxcenter, Ba583a, Erasmus MC, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Netherlands*
E-mail p.w.j.c.serruys@erasmusmc.nl

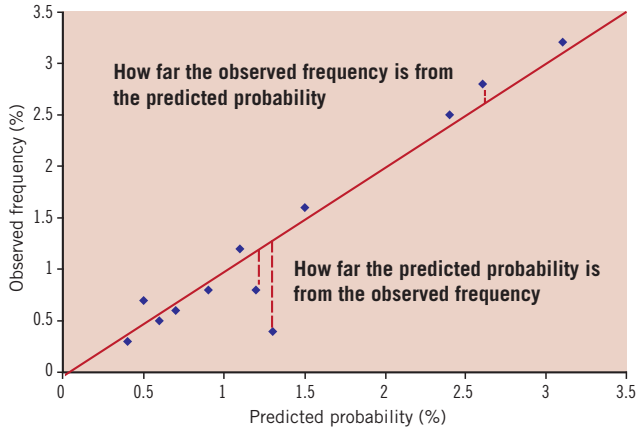

Figure 1. The concept of the Brier Score.

to assess the outcomes of each group. A well-discriminated model should therefore be able to discriminate between a trend towards a significantly different event rate within each respective category. For example, higher, intermediate and lower event rates should be discernible by their respective scores from the prediction models. Receiver operator characteristic (ROC) curves are commonly used to assess discrimination and are essentially a plot of true positive rate (sensitivity) of the score against false positive rate (1-specificity or 1-true negative rate). The area under the ROC curve (AUROC) gives an indication of the ability of the score to discriminate between those who do and do not experience the event with 0.5 being no better than chance and 1.0 a perfectly discriminated model.[3,5-8]

Conversely, calibration assesses how closely the predicted probabilities from the risk model agree with the actual outcomes (i.e., detecting a lack of goodness-of-fit). In keeping with the weather forecast analogy, this gives a probability of the forecast event and how close this prediction would be to the actual forecast event if and when it occurs. With the risk model however, this would be the agreement of all the predicted probabilities against their respective observed outcomes, which would give an indication as to how well calibrated our model was.

The Hosmer-Lemeshow goodness-of-fit test is frequently used to assess for calibration by assessing for the presence or absence of goodness-of-fit (based on chi-squared analysis and the subsequent significance of the *p*-value) for logistic regression models. A significant *p*-value means the overall model fit is NOT good, it however gives no indication of the nature of the goodness-of-fit. Within this test, observed outcomes are plotted by deciles of predictions, with a good discriminating model having more spread between such deciles compared to a poorly discriminating model. Good calibration and good discrimination are therefore usually inconsistent for predictive models, with a necessary trade-off between the two being required.[5-8]

Lastly is the proposition of potentially assessing whether the score will work in different populations from the population from which the score was derived. This can be performed with internal validation; i.e., performed on two separate samples within the study population, with the score derived in one sample and tested on the other as performed by Ito et al[11] in this issue of EuroIntervention.

Conversely, external validation is where the score is assessed on a separate population from the study group. The former would perhaps lead to a more optimistic assessment, and the latter, a potentially more accurate assessment of the validity of the score model. Other ways to cross-validate the models include methods such as "boot-strapping" and methods analogous to "jack-knifing," the former is described by Baran et al[10] in this issue of EuroIntervention they are however, outside the scope of this editorial.

Within this issue of EuroIntervention, three articles using these models are included.

Ito et al[11] developed a risk model to predict 30-day MACCE from the STENT Group Registry. The strengths of this study are that c-statistics, Hosmer-Lemeshow test of goodness-of-fit and an internal validation of the data were all performed. The latter was feasible given the large cohort of patients (>10,000 patients) investigated. The final c-statistic value was moderate (0.653 and 0.692 in the study and validation set) and was used by the authors to compare this model against other previously investigated models, despite the limitations of using c-statistics alone in comparing models as previously discussed.

Baran et al[10] developed a risk model for the VLST risk score for the second year post-DES implantation. Once again c-statistics and Hosmer-Lemeshow tests were performed; in this case a bootstrap method was used as a validation tool. Given the expected low event rate associated with stent thrombosis and moderate population size (approximately 7,500 patients), which is essentially underpowered to fully investigate stent thrombosis, it is noteworthy to see that a risk model could still be developed and does hold out the intriguing possibility of developing a model with a better discriminatory value within a larger patient group. The limitations, such as only being performed with one DES type and not comparing with BMS are obvious. However, the potential clinical utility with regards to this model are yet to be explored, and does potentially open the door with regards to perhaps better advising patients with respect to dual antiplatelet therapy regimes and even possibly the selection of PCI techniques, despite the studies limitations.

Federspiel et al[12] describe the fascinating concept of risk-benefit trade-off in the choice of coronary revascularisation modality: essentially trading the long-term risk of repeat revascularisation in exchange for short-term morbidity benefits. This issue is particularly pertinent in this present age, where the need for some individuals to remain active in their professional/personal lives are vital, and they are thus prepared to accept the longer term risks of coronary revascularisation in order to remain at their present functional state. Although this study was performed on the original ARTS study[13] data, which in itself was undertaken over ten years previously, the results are nevertheless supportive of this concept, and do allow a quantification of a level of risk that a patient would be able to accept in order to maintain their present state. Clearly, and as to what the authors elude to in their discussion, in order to better calculate the risk, data from SYNTAX[14] and FREEDOM trials, will be able to explore this concept to match modern day practice. This present study, however, is a welcome addition to the data in helping to explain these complex concepts to patients, and gives a taste of the quality of the

data due to come from further studies investigating this issue. In closing, it will be interesting to see how far we should go in allowing assessment of risk scores and importantly, allowing comparison of different types of risk models, within cardiology based trials. Undoubtedly, a greater collaboration with statisticians with expertise in these fields and cardiologists would aid in developing, refining and simplifying the assessment of these performance models.

References

1. Capodanno D, Capranzano P, Di Salvo ME, Caggegi A, Tomasello D, Cincotta G, Miano M, Patane M, Tamburino C, Tolaro S, Patane L, Calafiore AM. Usefulness of SYNTAX score to select patients with left main coronary artery disease to be treated with coronary artery bypass graft. *JACC Cardiovasc Interv.* 2009;2:731-738.

2. Capodanno D: Merging Anatomic and Clinical Risk Scores. Transcatheter Cardiovascular Therapeutics, Washington DC, US, September 21-25, 2010.

3. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21:128-38.

4. Cook NR. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation.* 2007;115:928-935.

5. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ.* 2009;338:b606.

6. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ.* 2009;338:b605.

7. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ.* 2009;338:b604.

8. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ.* 2009;338:b375.

9. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly weather review.* 1950;78:1-3.

10. Baran KW, Lasala JM, Cox DA, Mascioli SR, Song A, Deshpande MC, Jacoski MV, Dawkins KD. A clinical risk score for prediction of very late stent thrombosis in drug eluting stent patients. *EuroIntervention.* 2011;6:949-954.

11. Ito H, Nussbaum M, Hermiller JB, Hodes Z, Brodie B, Cheek B, Juk S, Krainin F, Metzger C, Duffy P, Humphrey A, Laurent S, Simonton CA. An integer based risk score for predicting 30-day major adverse cardiac or cerebrovascular events after percutaneous coronary intervention with drug-eluting stents: results from a large prospective multicentre registry, the STENT Group. *EuroIntervention.* 2011;6:942-948.

12. Federspiel JJ, Stearns SC, van Domburg RT, Sheridan BC, Lund JL, Serruys PW. Risk-benefit trade-offs in revascularisation choices. *EuroIntervention.* 2011;6:936-941.

13. Serruys PW, Ong AT, van Herwerden LA, Sousa JE, Jatene A, Bonnier JJ, Schonberger JP, Buller N, Bonser R, Disco C, Backx B, Hugenholtz PG, Firth BG, Unger F. Five-year outcomes after coronary stenting versus bypass surgery for the treatment of multivessel disease: the final analysis of the Arterial Revascularization Therapies Study (ARTS) randomized trial. *J Am Coll Cardiol.* 2005;46:575-581.

14. Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, Stahle E, Feldman TE, van den Brand M, Bass EJ, Van Dyck N, Leadley K, Dawkins KD, Mohr FW. Percutaneous coronary intervention versus coronary-artery bypass grafting for severe coronary artery disease. *N Engl J Med.* 2009;360:961-972.