ROC curves for clinical prediction models part 3. The ROC plot: a picture that needs a 1000 words

Calster, B. van; Wynants, L.; Collins, G.S.; Verbakel, J.Y.; Steyerberg, E.W.

Journal of Clinical Epidemiology 126 (2020) 220–223

Journal of Clinical Epidemiology

ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES

ROC curves for clinical prediction models part 3. The ROC plot: a picture that needs a 1000 words

Ben Van Calster[a,b,c,*], Laure Wynants[a,d], Gary S. Collins[e,f], Jan Y. Verbakel[c,g,h], Ewout W. Steyerberg[b]

[a]KU Leuven, Department of Development and Regeneration, Leuven, Belgium
[b]Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands
[c]EPI-Centre, KU Leuven, Leuven, Belgium
[d]Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, the Netherlands
[e]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Musculoskeletal Sciences, University of Oxford, Oxford, UK
[f]NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK
[g]Academic Centre for Primary Care, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium
[h]Nuffield Department of Primary Care Health Sciences, University of Oxford, UK

Accepted 24 May 2020; Published online 18 June 2020

In our recent perspective, we argue that receiver operating characteristic (ROC) curves add no useful information to the area under the ROC curve (AUC) for the evaluation of discriminatory ability of risk prediction models [1]. Our key argument is that a risk threshold needs to be considered if a model is used to support decision-making, while ROC curves in their standard form suppress threshold information. Despite her critical assessment of our article, Janssens raised no strong or convincing arguments against this view [2]. Janssens states that the ROC curve is valid, and that it is subjective whether it has added value over the AUC alone: ''others may benefit from seeing how and how much (little) the addition of predictors improves the discriminative ability''. However, that is exactly what the AUC quantifies, and Janssens seemingly agrees: ''the AUC and ROC plot present the overall discriminative ability of prediction models''.

1. The threshold concept

We discussed various limitations of ROC curves, but in this response, we focus on the threshold issue. We write ''the performance of risk prediction models for decision-making has to be conditional on a risk threshold to fix misclassification costs'' [1]. The risk threshold is related to how false negatives are valued in relation to false positives. Let us consider the following scenario for a patient treated with chemotherapy for metastatic nonseminomatous testicular cancer. Residual disease can be surgically resected through lymphadenectomy. If physicians agree that only patients with a risk of 20% and higher should have lymphadenectomy, they are more worried about false negatives than about false positives. They do not want to avoid false negatives at any cost. Operating on patients with a risk of 20% yields 4 false positives to prevent 1 false negative. Janssens wonders whether this 1:4 benefit-to-harm ratio ''applies to the threshold or to the entire group that is selected by the threshold'' [2]. Among patients with a risk $\geq$20%, the ratio of diseased versus nondiseased patients will be less than 1:4; hence, Janssens wonders whether a 1:4 benefit-to-harm ratio requires a threshold below 20%. It does not. If a risk of 20% is the lower boundary for surgery, up to 4 false positives are acceptable to avoid 1 false negative: the benefit-to-harm ratio applies to the threshold. If the 1:4 ratio would apply to the group selected by the threshold, we would need to treat everyone when the event rate of residual disease in the population is $\geq$ 20%. It would also imply that the threshold for treatment is lower for models with higher AUC.

In accordance with our perspective, we agree with Janssens that comparing models cannot be based on ROC or AUC alone. For example, a model may have a high AUC, but may systematically overestimate risk in all patients [3]. Model comparison also requires assessing calibration (the reliability of risk estimates) and potential clinical utility for decision-making (e.g., using net benefit and decision curve analysis). The net benefit quantifies the utility of decision-making for a given threshold and the associated benefit-to-harm ratio [4]. The decision curve plots the net benefit for a range of clinically sensible thresholds because there is often no threshold that applies to all settings. Note that the choice of a threshold is a clinical rather than a statistical decision [5].

* Corresponding author. Tel.: +32 16377788; fax: +32 16344205.
E-mail address: ben.vancalster@kuleuven.be (B. Van Calster).

https://doi.org/10.1016/j.jclinepi.2020.05.037

2. Example

Let us assume that 55% of patients have residual disease (the event), and that patients with a risk ≥20% have surgery. We compare two models. Both include the same continuous predictor but a different—uncorrelated—binary predictor. Model A includes a binary predictor with a sensitivity of 88%, a specificity of 49%, and prevalence of 71%. Model B includes a predictor with a sensitivity of 52%, a specificity of 93%, and prevalence of 32%. Model A and B have AUCs of 0.78 and 0.82, respectively. We validate the model on data from 100,000 patients (Fig. 1), and the models are well calibrated. The ROC curves have a different shape, as expected. For model A, the curve is higher in the top right and for model B, in the lower left. This suggests that model A is preferable at low thresholds. However, at the 20% risk threshold, model B actually has higher sensitivity (97.0% vs 96.7%), but lower specificity (19.1% vs 26.4%). This can be explained by the bimodal risk distribution for events with model B. What does this mean for decision-making? At the 20% threshold, model A has highest net benefit, despite lower sensitivity and overall AUC than model B.

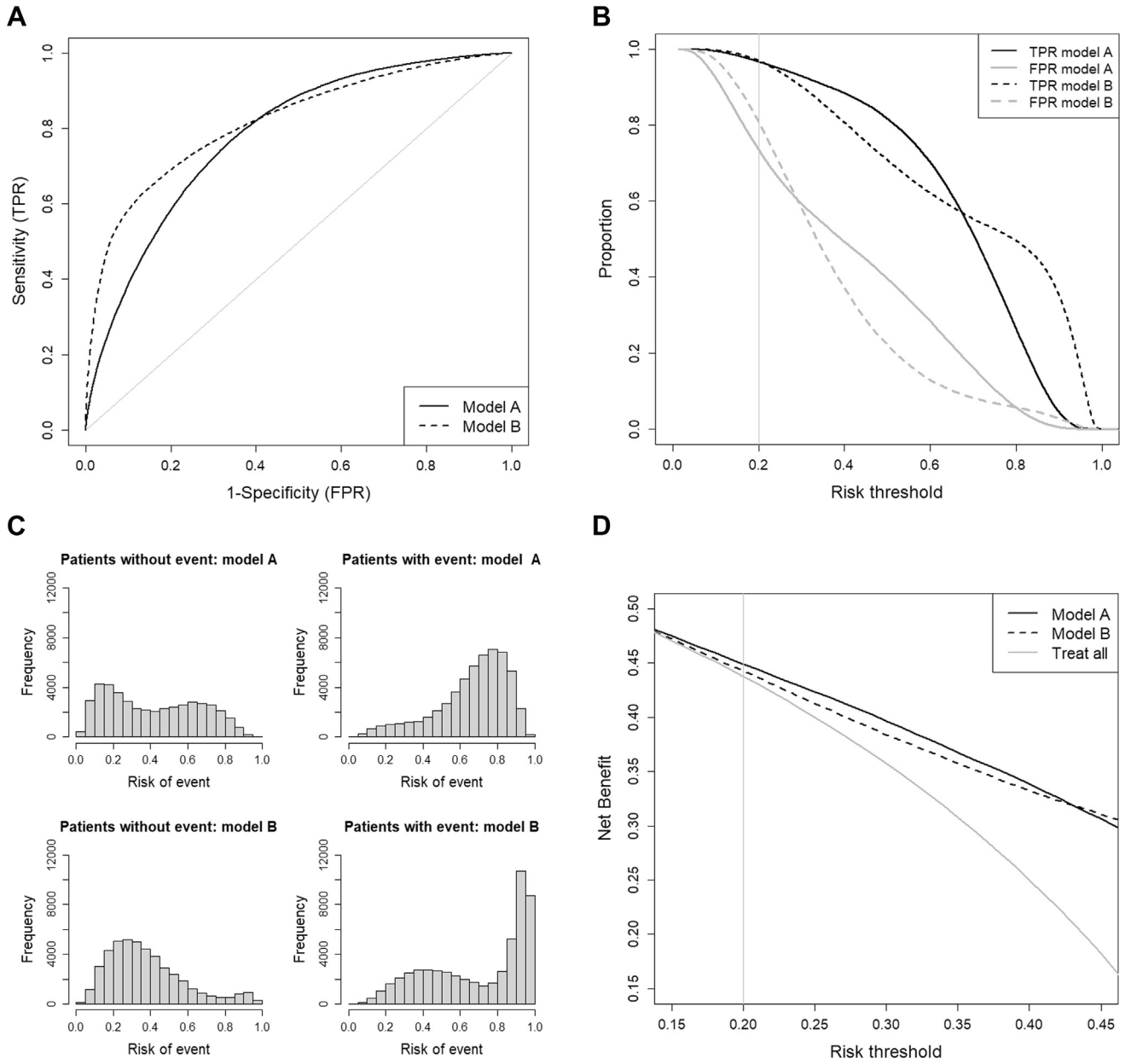

Fig. 1. The ROC curves (A), classification plots (B), risk distributions (C), and decision curves (D). *Abbreviations:* ROC, receiver operating characteristic.

Now, suppose model A was miscalibrated. The model was developed in a population with a much lower event rate, leading to clear underestimation of the risk of event in our population. Underestimation of risk (i.e., an incorrect intercept) does not affect the ROC curve (Fig. 2). However, misclassification plots and histograms have dramatically changed for model A. As a result, the decision curve indicates that model B now has higher net benefit at the 20% threshold. The impact of miscalibration was not captured at all by the ROC curve.

Two other points raised by Janssens also merit attention. First, we do not state that different ROC curve shapes for models with the same AUC are the result of random fluctuation. Rather, when validating a model on two equally large samples from the same population, the ROC curve can be very different for the same AUC. In particular with limited sample size, ROC curves have unstable shapes. Second, the suggestion that ROC curves "hint at sample size, the percentage of cases, or the number of different risk estimates" is crude. The sample size should be determined a priori and

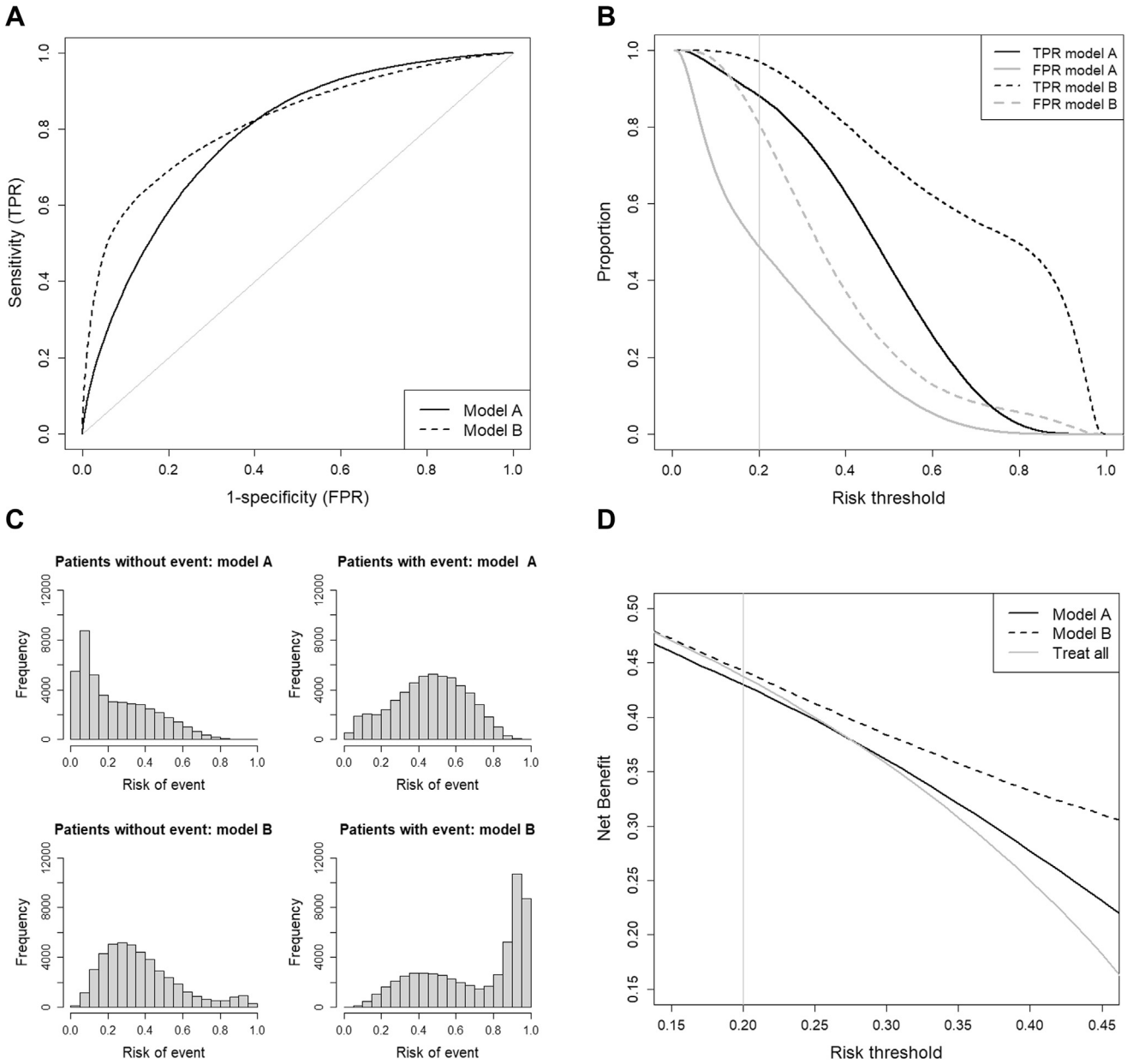

Fig. 2. The ROC curves (A), classification plots (B), risk distributions (C), and decision curves (D) when the model with binary predictor A is miscalibrated (risks strongly underestimated). *Abbreviations:* ROC, receiver operating characteristic.

reported together with the event rate, as recommended in the TRIPOD statement [6].

To conclude, the AUC quantifies discriminatory ability, and ROC curves do not add interpretable information to the AUC, particularly when thresholds are omitted. Calibration and decision curves are pivotal [7]. To visualize discrimination, classification plots or risk distributions are more informative than ROC curves without thresholds, which remain a waste of space in medical journals.

References

[1] Verbakel JY, Steyerberg EW, Uno H, De Cock B, Wynants L, Collins GS, et al. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. J Clin Epidemiol 2020.

[2] Janssens ACJW. The ROC plot: the picture that could be worth a 1000 words. J Clin Epidemiol 2020.

[3] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17:230.

[4] Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and interpreting decision curve analysis: a guide for investigators. Eur Urol 2018;74:796−804.

[5] Wynants L, van Smeden M, McLernon D, Timmerman D, Steyerberg EW, Van Calster B. Three myths about risk thresholds for prediction models. BMC Med 2019;17:192.

[6] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol 2015;68:134−43.

[7] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128−38.