ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models

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

ROC CURVES FOR CLINICAL PREDICTION MODEL SERIES

ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models

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

[a]*KU Leuven, Department of Public Health and Primary Care, Leuven, Belgium*
[b]*Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK*
[c]*Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands*
[d]*Division of Population Sciences, Dana-Farber Cancer Institute, Boston, MA, USA*
[e]*KU Leuven, Department of Development and Regeneration, Leuven, Belgium*
[f]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[g]*Oxford University Hospitals NHS Foundation Trust, Oxford, UK*

Abstract

Objectives: Receiver operating characteristic (ROC) curves show how well a risk prediction model discriminates between patients with and without a condition. We aim to investigate how ROC curves are presented in the literature and discuss and illustrate their potential limitations.

Study Design and Setting: We conducted a pragmatic literature review of contemporary publications that externally validated clinical prediction models. We illustrated limitations of ROC curves using a testicular cancer case study and simulated data.

Results: Of 86 identified prediction modeling studies, 52 (60%) presented ROC curves without thresholds and one (1%) presented an ROC curve with only a few thresholds. We illustrate that ROC curves in their standard form withhold threshold information have an unstable shape even for the same area under the curve (AUC) and are problematic for comparing model performance conditional on threshold. We compare ROC curves with classification plots, which show sensitivity and specificity conditional on risk thresholds.

Conclusion: ROC curves do not offer more information than the AUC to indicate discriminative ability. To assess the model's performance for decision-making, results should be provided conditional on risk thresholds. Therefore, if discriminatory ability must be visualized, classification plots are attractive. © 2020 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Receiver operating characteristic curve; Risk prediction models; Classification plots; Risk threshold

1. Introduction

Clinical risk prediction models to predict the risk of having (diagnostic research) or developing (prognostic research) an event are being published in increasing numbers. These models provide individual risk predictions based on the values of the predictors included in the model [1]. Evaluating predictive performance is an important step in determining a model's potential usefulness. One key aspect of performance is discrimination, the ability to separate those with and without an event by

Authors' contribution: B.V.C. conceived the article content and structure, with initial feedback from E.S., G.S.C., and J.Y.V. H.U. developed most of the source code, and B.D.C. and J.Y.V. developed the R function. J.Y.V. applied the methods to the examples. B.V.C. and J.Y.V. wrote the first draft. All authors contributed to the article revision, which included adding new text pertinent to their expertise and refining the examples for the intended audience. B.V.C. is the guarantor.

* Corresponding author. KU Leuven, Department of Development and Regeneration, Herestraat 49 Box 805 3000, Leuven, Belgium. Tel.: 0032 16 377788; fax: 0032 16 344205.

E-mail address: ben.vancalster@kuleuven.be (B. Van Calster).

What is new?

Key findings
- Published studies evaluating the performance of a clinical prediction model commonly present receiver operating characteristic (ROC) curves without including threshold information.

- ROC curves have an unstable shape and can result in problematic and potentially misleading model comparison by a risk threshold.

What this adds to what was known?
- ROC curves add little information on discriminatory ability to the area under the curve (AUC), particularly when the curve does not include thresholds.

What is the implication and what should change now?
- We recommend focusing on the AUC to indicate discriminative ability and reporting sensitivity and specificity at clinically relevant thresholds to report the potential of the model for decision-making.

- When a visualization of performance for decision-making is desired, we argue that classification plots are more informative than ROC curves, as they show sensitivity and specificity conditional on risk thresholds.

predicting higher risks for patients with the event than patients without.

Discrimination performance is often visualized using a receiver operating characteristic (ROC) curve [2−4]. ROC curves show classification performance conditional on consecutive thresholds, but do not show the thresholds themselves. The ROC curve can be summarized by the area under the curve (AUC) to quantify discrimination. The AUC estimates the probability that a model can correctly discriminate between randomly selected individuals with and without the event.

The aim of a prediction model is typically to identify patients who have sufficiently increased risk to warrant receiving a given treatment or intervention [5]. A risk threshold is thus specified. Although thresholds are often chosen in a pragmatic data-driven way based on a desired combination of sensitivity and specificity, the threshold itself has a decision-analytic meaning. The risk threshold has a direct relationship with the relative cost of false positives and false negatives [6]. For example, if we decide

that patients should be treated when their risk of an event is 10% (odds 1:9), our decision implies that correctly treating one patient with the event justifies unnecessarily treating at most nine patients without the event. Choosing a clinically relevant risk threshold depends on the context, such as the target population and specific treatment or intervention.

As ROC plots in their standard form do not show thresholds, they cannot be used to assess model performance conditional on the risk threshold. This severely limits the interpretability of ROC curves. We therefore argue that ROC curves in their standard form add little information to the AUC, as was highlighted in the TRIPOD Explanation and Elaboration document [7].

In this study, we investigate whether and how ROC curves are presented in peer-reviewed publications that externally validate clinical risk prediction models. We assess and illustrate the limitations of ROC curves using a case study and simulated data. We compare the ROC curve with an alternative, the classification plot.

2. Methods

2.1. The ROC curve and AUC

To use a multivariable risk prediction model to select patients for an intervention, a classification threshold needs to be specified using the estimated risk. Risks below the risk threshold imply test negative (or low risk) and above the threshold imply test positive (or high risk) (Table 1). In practice, a low threshold results in high sensitivity (high true positive rate) and low specificity (high false positive rate). A high threshold results in low sensitivity and high specificity. Lower risk thresholds imply lower relative costs of false positives vs. false negatives [6].

ROC curves display the true positive rate (y-axis) against the false positive rate (x-axis) for each possible threshold. The better the model discriminates, the more the ROC curve approaches the upper left corner of the plot. A model with no discriminative ability has a true ROC curve that lies on the diagonal line.

The AUC quantifies the model's discriminative ability. For dichotomous outcomes, the AUC is equivalent to the concordance probability or 'c-statistic', which is the probability that the model estimates higher risks for patients with the event than patients without the event [8−10]. A model that perfectly discriminates between patients with and without an event would have an AUC of 1 (the theoretical maximum), whereas a model with no ability to discriminate between such patients would have an AUC of 0.5. The TRIPOD reporting guideline for clinical prediction models recommends reporting the AUC and its confidence interval (CI) [11].

A convenient method for obtaining a CI for the AUC is the logit transform method, which produces asymmetric

Table 1. Definitions of terms used in this study

Term	Explanation
(Risk) threshold	Value of the estimated risk that is used to classify patients as test positive (increased risk of the event or target condition) or test negative.
Sensitivity	Proportion/percentage of patients with the event or target condition that are classified as test positive. Also called the true positive rate.
Specificity	Proportion/percentage of patients without the event or target condition that are classified as test negative. Also called the true negative rate.
False positive rate	Proportion/percentage of patients without the event or target condition that are classified as test positive. Calculated as 1—specificity.
AUC	The area under the ROC curve. For binary outcomes (event vs. no event), the AUC equals the c-statistic, which is the probability that a random patient with the event has a higher estimated risk than a random patient without the event or target condition.

intervals that do not exceed 1 [12,13]. Both the time horizon and the presence of censored observations should be considered for time-to-event data [14,15]. Naive calculations of true and false positive rates for a chosen threshold can be misleading as some of the censored observations would have had events if follow-up was longer. It is feasible to construct ROC curves and calculate AUCs/c-statistics that are conditional on follow-up time [16], but this might further complicate interpretation [14,15,17−19].

2.2. Classification plots

Classification plots can also be used to visualize discrimination. They show the true and false positive rates by risk thresholds, explicitly allowing the reader to evaluate how the sensitivity and false positive rate (or specificity) change with thresholds. Examples of such plots already exist in the literature [20]. Fig. 1 shows hypothetical classification plots for various AUCs and event rate scenarios (i.e., the outcome prevalence or incidence, as appropriate).

2.3. ROC curves in the literature

To gain insight into commonly used methods in the literature, we performed a full-text search of external validation studies of clinical prediction models based on logistic regression for binary outcomes. We ran a pragmatic literature search in MEDLINE (July 12, 2018) to identify studies that mentioned "external validation" and "model" in the title. We extracted data on the use of ROC curves and the AUC (Table 2; Appendix Table A.1).

2.4. Illustrative case study

We illustrate the limitations of ROC curves using a case study. We used data from a previous study on patients with metastatic nonseminomatous testicular cancer who were treated with chemotherapy [21,22]. After chemotherapy, retroperitoneal lymph nodes can either still contain remnants of the metastases (mature teratoma or viable cancer) or only contain benign necrosis. Surgical lymph node resection can provide certainty and remove metastatic remnants,

but should be avoided when lymph nodes only contain benign necrosis.

We developed and validated a clinical prediction model to predict the presence of metastatic remnants, using malignancy as a binary outcome. We used data from 544 patients to develop the model and data from 550 patients to externally validate the model [22]. Malignancy was present in 299 (55%) patients in the development data set and in 370 (67%) patients in the validation data set. We built a baseline model with three predictors: the maximal diameter of the residual mass, percentage reduction in mass size after chemotherapy, and presence of teratoma elements in the primary tumor. We also built an extended model, which added the presence of elevated alpha-fetoprotein levels and elevated human chorionic gonadotropin levels to the baseline model.

2.5. Simulated data

We used two sets of simulated data to illustrate the characteristics and limitations of ROC curves.

Simulation setting A

In the first setting, we considered a binary outcome with an event rate of 22%: the event is observed in 22% of patients and not observed in the remaining 78%. The outcome was predicted through a predictor with an underlying normal distribution (mean 1, variance 1) that had a true AUC of 0.74. We simulated data sets with sample sizes of 50 (11 events on average), 100 (22 events), 250 (55 events), and 500 (110 events). We used the data to investigate the variability and stability of ROC curves.

Simulation setting B

One ROC curve completely 'dominates' another when it has a higher AUC, and the ROC curves do not ever cross. The dominant model has a higher sensitivity at every specificity level and a higher specificity at every sensitivity level than the other model. However, the dominant model may not have both higher sensitivity and higher specificity when the same risk threshold is used for both models [23]. In practice, the dominant model often has lower sensitivity
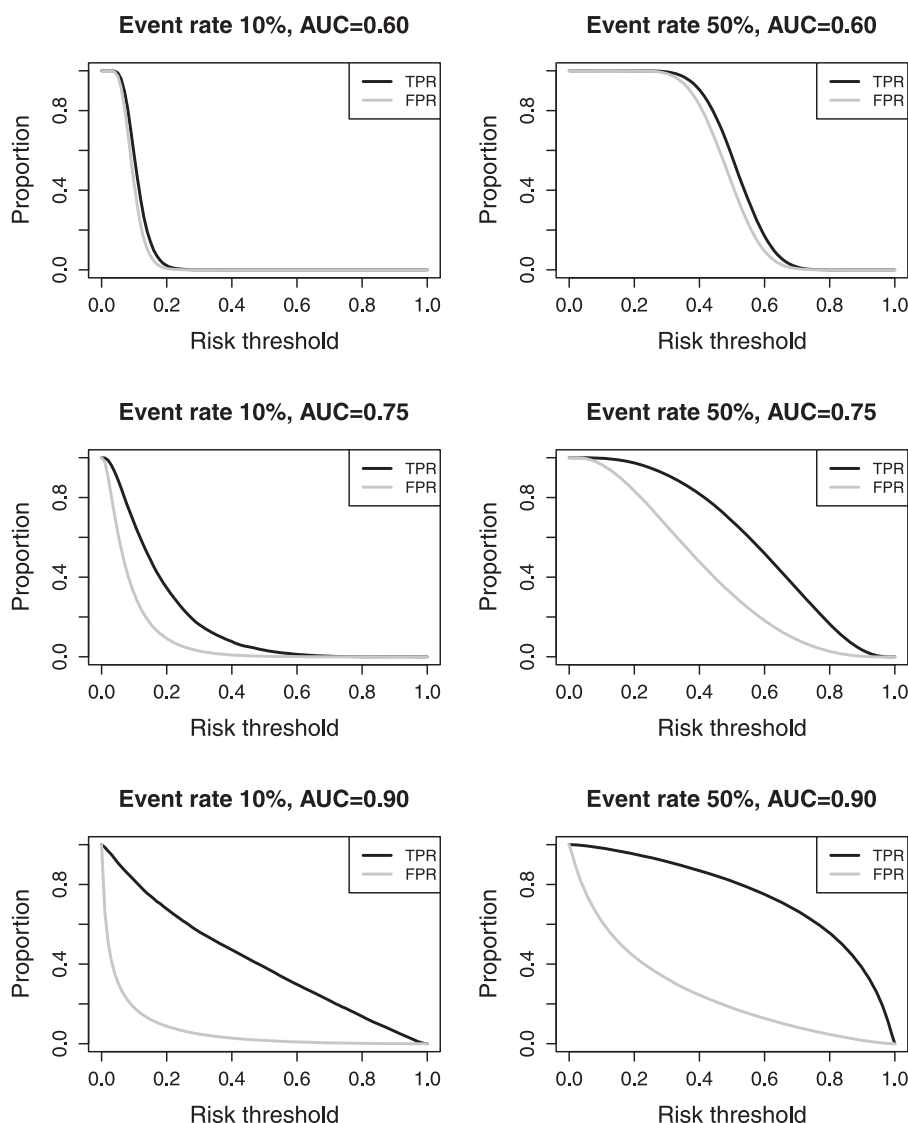
Fig. 1. Theoretical classification plots for scenarios determined by AUCs (0.6, 0.75, or 0.9) and outcome event rate (10% or 50%). The curves assume normally distributed linear predictors.

and higher specificity (or vice versa) than the other model at certain risk thresholds [23]. Although counterintuitive, the decrease in one metric is compensated for by the increase in the other metric, thus increasing the overall discriminatory ability of the model with higher AUC. ROC curves do not expose this counterintuitive behavior. Even when risk thresholds are indicated on the curves, this issue is not easily appreciated.

To illustrate this ROC curve limitation, we simulated a data set of 100,000 patients, of which 50,000 had the event (event rate 50%). We assumed two uncorrelated predictive markers. Both markers were normally distributed with mean 0 and variance 1 in patients without the event, and normally distributed with mean 0.8 and variance 1 in patients with the event. We compared a model with one

marker (true AUC = 0.71) to a model with both markers (true AUC = 0.79).

3. Results

3.1. Literature search

The pragmatic literature search identified 86 published studies that externally validated a clinical prediction model. Thirty-three of the studies (38%) did not present ROC curves, of which six also did not report the AUC (Table 2). Fifty-two of the studies (60%) presented ROC curves without a threshold on the curve. In one of these 52 studies, the authors presented a 'binary' ROC curve of dichotomized predictions after applying one risk threshold.

Table 2. Results of the literature search

Search: Studies that externally validate a logistic regression model	
Studies included in our assessment	86
Publication year	
2002–2010	14 (16%)
2011–2015	26 (30%)
2016	17 (20%)
2017	18 (21%)
2018	11 (13%)
Evaluation of the use of ROC curves and AUCs	
The ROC curve without thresholds, with AUC	52 (60%)
The ROC curve with multiple thresholds shown, with AUC	1 (1%)
No ROC curve or AUC	6 (7%)
No ROC curve, with AUC	27 (31%)

Only one study (1%) presented ROC curves with multiple thresholds.

3.2. Case study: single model

Using case study data, we constructed and externally validated a baseline model of three predictors to predict the presence of metastatic remnants of testicular cancer, using malignancy as a binary outcome. The AUC of the externally validated baseline model was 0.77 (95% CI 0.73 to 0.81). Fig. 2A shows the accompanying ROC curve. Using a risk threshold of 0.2 (20%), previously suggested as clinically reasonable, resulted in a sensitivity of 99% (95% CI 97 to >99%) and specificity of 7% (95% CI 3 to 10%) [21]. We selected three risk thresholds (0.2, 0.3, and 0.4) to indicate on the ROC curve (Fig. 2B).

Fig. 3A shows the classification plot for the same model, including 95% CIs for the selected thresholds. The sensitivity and false positive rate for each threshold value are shown below the plot. Other statistics could also be displayed, such as the positive and negative predictive values or the diagnostic odds ratio.

3.3. Case study: model comparison

ROC curves are often used to compare the discrimination performance of clinical prediction models [24]. Fig. 2C shows ROC curves for the baseline and extended testicular cancer models. To compare discrimination, we can compute a CI for the difference in c-statistics (AUCs) using DeLong's method, although the method is biased in certain circumstances [12,25,26]. The AUC of the externally validated extended testicular cancer model was 0.80 (95% CI 0.77 to 0.84). The difference in AUC between the two models was 0.03 (95% CI −0.01 to 0.08).

Sensitivity and specificity are usually compared at one or more thresholds, using the same risk thresholds for both models. At a 20% risk threshold, the extended testicular cancer model had a sensitivity of 99% (95% CI 97% to 100%) and a specificity of 16% (10% to 21%). The extended model had 0.3% (95% CI −0.9% to 1.5%) greater sensitivity and 8.9% (95% CI 4.2% to 13.6%) greater specificity than the baseline model.

As ROC curves do not align clinical prediction models by risk thresholds, they do not allow direct comparison by thresholds [27]. We can show the thresholds on the curves (Fig. 2D), but the plot quickly becomes cluttered as more thresholds are added.

Classification plots can show comparative results for two clinical prediction models (Fig. 3B), allowing performance to be directly compared by thresholds. Detailed results, including CIs for selected thresholds, can be shown in a table. Software to create classification plots is available, for example, in the ROCR [28] and rmda [20] packages for R. We have also developed an R function to produce classification plots, allowing the user to select a range of performance statistics (plotted on the curve and/or printed below the plot) (Appendix file B.1 and available on GitHub via https://github.com/janverbakel/ClassificationPlot). The function allows users to plot pointwise 95% CIs at their thresholds of interest.

3.4. Simulated data: variability of ROC curves

Fig. 4 shows how the ROC curves from simulation setting A vary with sample size. The variability was enormous for $n = 50$ and remained considerable even at $n = 500$. Although uncertainty is inevitable in any data set, ROC curves have a source of variation that the AUC does not. Two curves for the same prediction model evaluated on two samples from the same target population can have the same AUC, but very different shapes. This adds complexity without adding useful information.
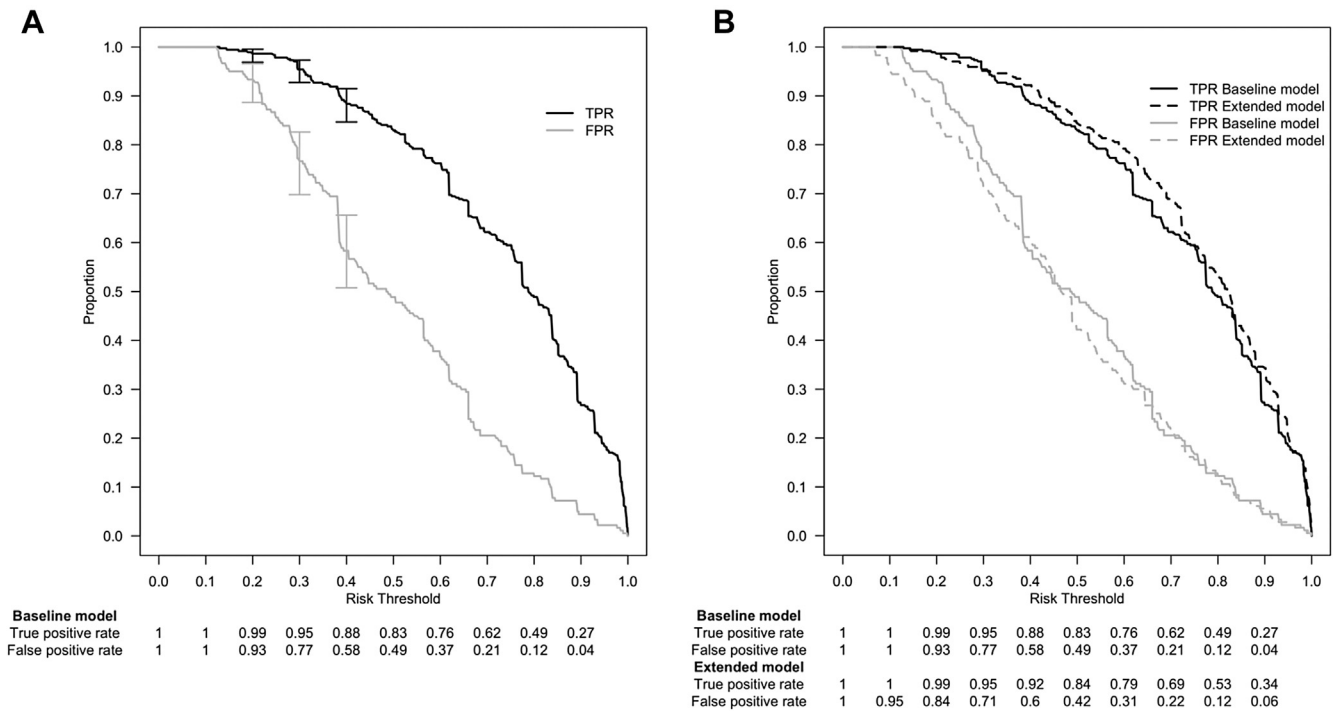
To illustrate the issue, Appendix Fig. A1 repeats Fig. 4 but shows 100 curves with an AUC of 0.75. The ROC curves vary considerably even with a sample size of 500, implying that we require very large samples to obtain stable curves. Appendix Fig. A2 shows the ROC curve for the baseline model from the case study with a 95% confidence band for the curve based on Martínez-Camblor's method [29] and 95% CIs for sensitivity and specificity at selected thresholds. The confidence band for the curve is wider than the CIs for the thresholds of interest. Confidence bands for the whole ROC curve integrate the uncertainty of sensitivity and specificity at every possible threshold, which adds up and causes wide bands.

Classification plots also suffer from sampling variability in their shape. To illustrate, Appendix Fig. A3 shows the classification plots for the ROC curves from Fig. 4. However, we are usually not interested in the full curve. As deciding whether a patient should receive an intervention is a binary decision, we are usually interested in the pointwise CIs for one or more clinically relevant thresholds (cf Fig. 3).
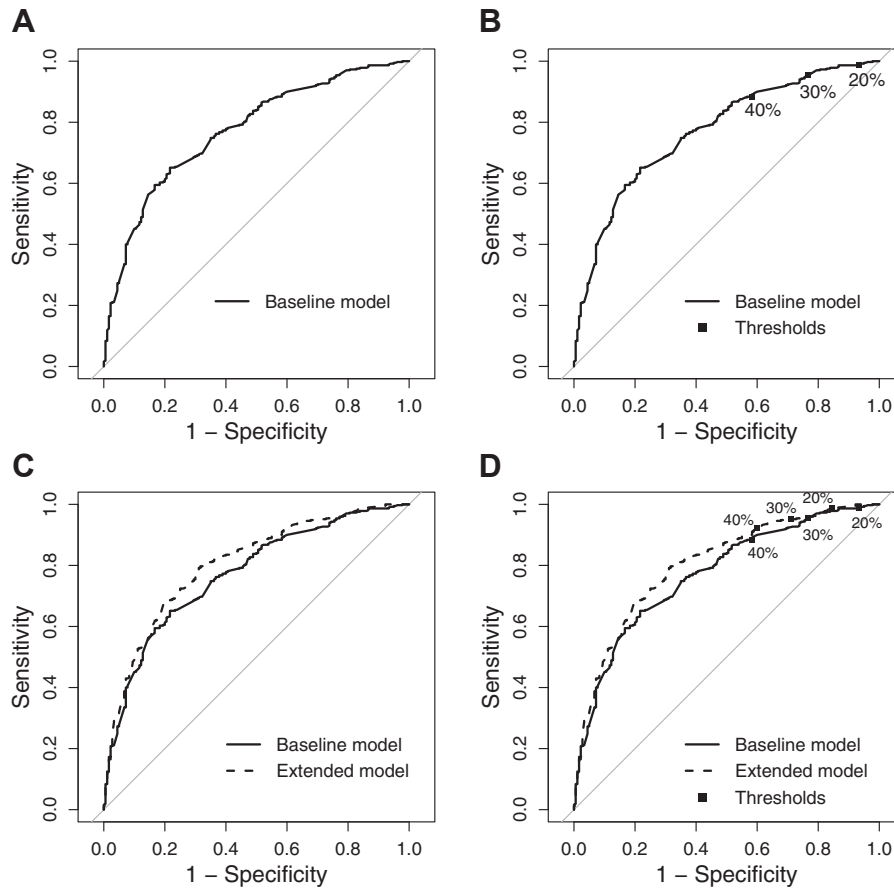
Fig. 2. Visualizing the discriminatory ability of the baseline and extended models for testicular cancer using ROC curves: (A and B) baseline model alone or (C and D) comparing the baseline and extended models, (A and C) without and (B and D) with risk thresholds.

Baseline model										
True positive rate	1	1	0.99	0.95	0.88	0.83	0.76	0.62	0.49	0.27
False positive rate	1	1	0.93	0.77	0.58	0.49	0.37	0.21	0.12	0.04

Baseline model										
True positive rate	1	1	0.99	0.95	0.88	0.83	0.76	0.62	0.49	0.27
False positive rate	1	1	0.93	0.77	0.58	0.49	0.37	0.21	0.12	0.04
Extended model										
True positive rate	1	1	0.99	0.95	0.92	0.84	0.79	0.69	0.53	0.34
False positive rate	1	0.95	0.84	0.71	0.6	0.42	0.31	0.22	0.12	0.06

Fig. 3. Classification plots showing sensitivity (true positive rate, TPR) and false positive rate (FPR) by threshold for (A) the baseline testicular cancer model (with 95% pointwise confidence intervals of TPR and FPR for risk thresholds of 20%, 30%, 40%) or (B) both the baseline and extended testicular cancer models.
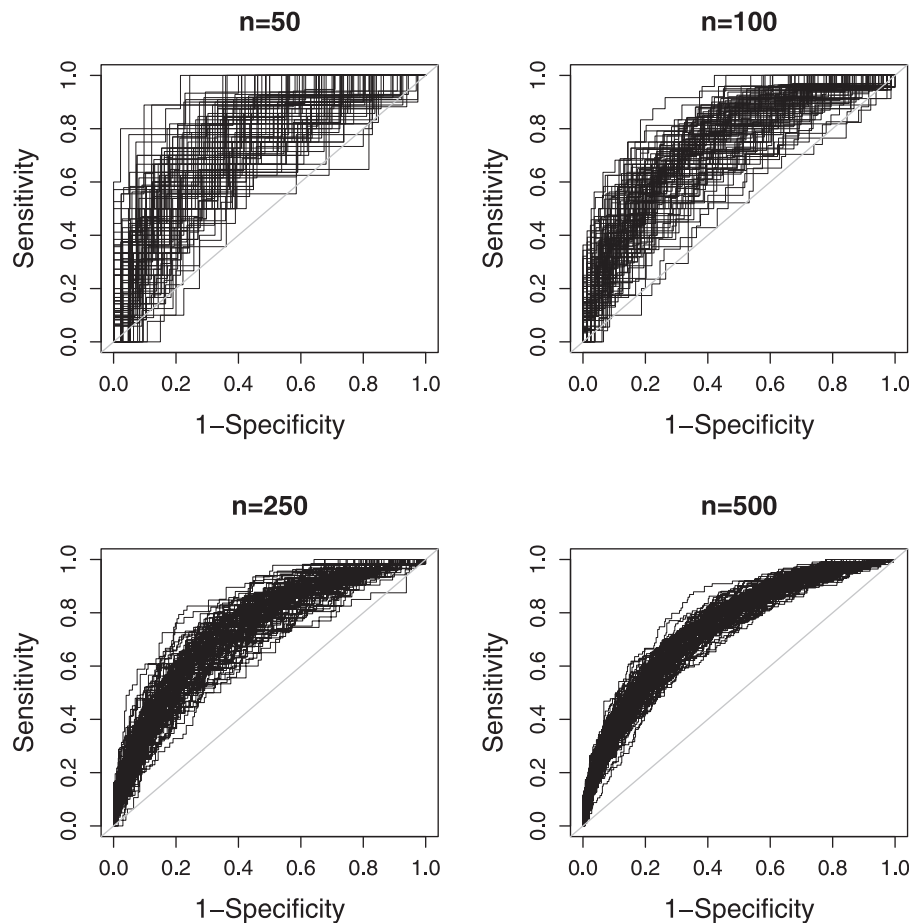
Fig. 4. Effect of increasing sample size (*n*) on the variability and shape of the ROC curve. We considered a normally distributed risk prediction model with a true AUC of 0.74 for predicting the presence of an event with a true event rate of 22%. For each of four sample sizes, 100 data sets were simulated and their corresponding ROC curves plotted. At lower sample sizes, ROC curves are more rough and variable than at high sample sizes.

3.5. Simulated data: model comparison can be misleading using ROC curves

Two models were generated for simulation setting B. The model with two markers had an ROC curve that was dominant over the ROC curve from the model with one marker (Fig. A2). However, the classification plot shows that the dominant model did not have a better sensitivity and false positive rate at every threshold (Fig. A4).

Comparing the two models developed using the case study data set shows a similar result (Fig. 3). However, that data set had a smaller sample size and hence relatively more random variation.

4. Discussion

We have shown that published studies about prediction models commonly present ROC curves without adding threshold information. We argued that ROC curves without thresholds provide no more information than the AUC if we are interested in the discriminative ability of a prediction model. We also illustrated several shortcomings of ROC curves: ROC curves (a) do not automatically show threshold information, (b) have highly unstable shapes (even when holding the AUC constant), (c) offer limited ability to compare two models conditional on thresholds, and (d) give potentially misleading results when comparing models where one ROC curve dominates the other.

The performance of risk prediction models for decision-making has to be conditional on a risk threshold to fix misclassification costs. Researchers must therefore present, at minimum, thresholds alongside ROC curves [27,30,31]. However, ROC curves remain inefficient even when a few thresholds are added to the plot. ROC curves only provide information for the added thresholds and do not have space for many thresholds. They also do not allow easy comparison of the performance of two models for the same threshold.

An alternative to the ROC curve is the classification plot, which explicitly shows classification performance conditional on thresholds. Admittedly, classification plots are

more complex because they require two lines instead of one, but offer the advantages of considering many thresholds and allowing easier evaluation of sensitivity and specificity by thresholds. Classification plots and ROC curves exhibit similar levels of variability, which is notably large in small samples. However, a single risk threshold is usually used when a model is used for decision-making in a specific setting. Although the particular threshold can vary by setting, a limited set or range of clinically relevant thresholds is usually of interest. Pointwise CIs are usually sufficient, instead of confidence bands for the full curve. Classification plots can easily be restricted to the range of thresholds of interest.

Our work should be interpreted within a wider context. Calibration and clinical utility are increasingly recognized as performance aspects that are at least as important as discrimination [32]. Calibration refers to the reliability of the risk estimates and can be poor even when discrimination is good. When externally validating a prediction model, we recommend using calibration curves to investigate whether the model is calibrated [33]. Clinical utility refers to a model's value for clinical decision-making. Common approaches for evaluating clinical utility are decision-curve analysis and relative utility [20,34−37]. These approaches avoid conducting a full-scale cost-effectiveness study. More detailed descriptions of these aspects of performance are beyond the scope of this study. However, we argue that visualizations of calibration and clinical utility are usually more relevant than visualizations of discrimination. When a risk threshold is used to identify high-risk patients from a model, calibration helps to determine whether the classification works as intended. For example, if risks are overestimated, the risk threshold identifies more high-risk patients than planned.

This study referred to the situation in which a novel marker is added to a prediction model. Specific measures and visualizations are available for this situation, which were beyond the scope of this study [30]. For example, research on added markers has focused on the potential increase in discrimination when the novel marker is added, rather than calibration or utility assessments [38]. In that context, it is worth mentioning that the classification plot has links with integrated discrimination improvement (IDI) [39]. The area under the sensitivity curve equals the mean predicted risk of event for events, whereas the area under the false positive rate curve equals the mean predicted risk of event for non-events. The difference between these areas is known as the discrimination slope [40]. If a new marker is added, the difference between discrimination slopes of the model with and without the marker is known as the IDI.

The study of biomarkers is a related area to that of multivariable prediction models. Although a detailed discussion of the value of ROC curves for biomarker studies is beyond the scope of this study, we mention a few considerations here. Biomarkers are often measured on different scales, making comparisons at the same threshold difficult. A comparison at the same risk threshold would be possible after modeling the marker on its own or in combination with other predictors. However, biomarkers are often investigated for their potential role in multivariable clinical prediction models [41]. In these situations, we suggest that the AUC alone is sufficient to summarize the biomarker's discriminative ability and that the difference in AUC is sufficient to compare biomarkers. Investigators can also evaluate a biomarker's value as a standalone criterion to identify patients who need additional testing. They will then be interested in the model's ability to classify a patient based on a threshold biomarker value. The choice of a threshold can be illustrated with ROC curves, even when taking misclassification costs into account [42,43]. For the actual derivation of an optimal threshold, however, an ROC curve is not useful. Many claim that an optimal threshold corresponds to the point on the ROC curve that is closest to the upper left corner [44,45]. However, this threshold is often suboptimal from a clinical perspective, as it does not consider clinically relevant costs of false positive and false negative classifications [46]. Alternatively, sensitivity is often compared at a fixed level of specificity, or vice versa. Such comparisons do not require an ROC curve. This approach implies the use of different risk thresholds for the two models, implying inconsistent definitions of a high-risk patient, which conflicts with rational decision-making [6].

To conclude, ROC curves are limited because they suppress threshold information and can greatly vary in their shape. Instead, we recommend focusing on the AUC to summarize discriminatory ability. When the impact on decision-making is of interest, performance at one or a few clinically relevant risk thresholds should be reported (with pointwise CIs). When discriminatory ability must be visualized, classification plots are attractive.

CRediT authorship contribution statement

Jan Y. Verbakel: Conceptualization, Software, Formal analysis, Writing - original draft, Visualization. **Ewout W. Steyerberg:** Conceptualization, Writing - review & editing, Supervision. **Hajime Uno:** Software, Writing - review & editing. **Bavo De Cock:** Software, Writing - review & editing. **Laure Wynants:** Conceptualization, Writing - review & editing. **Gary S. Collins:** Conceptualization, Visualization, Writing - review & editing. **Ben Van Calster:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Visualization, Supervision.

Acknowledgments

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.01.028.

References

[1] Van Calster B, Steyerberg EW, Harrell FH. Risk prediction for individuals. JAMA 2015;314:1875.

[2] Obuchowski NA, Lieber ML, Wians FH Jr. ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem 2004; 50:1118–25.

[3] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38(5):404–15.

[4] Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–35.

[5] Mallett S, Halligan S, Thompson M, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. BMJ 2012;345: e3999.

[6] Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med 1975;293:229–34.

[7] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73.

[8] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143: 29–36.

[9] Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: general issues and tail-area-based methods. Stat Med 2006;25:543–57.

[10] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 1975;12(4):387–415.

[11] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual Prognosis or diagnosis (TRIPOD). Ann Intern Med 2015;162:735–6.

[12] Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press; 2003.

[13] Qin G, Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Stat Methods Med Res 2008;17(2):207–21.

[14] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 2000; 56:337–44.

[15] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 2011;30:1105–17.

[16] Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 2017;17:53.

[17] Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA 1982;247:2543–6.

[18] Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61:92–105.

[19] Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. Stat Med 2006;25: 3474–86.

[20] Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. J Clin Oncol 2016;34:2534–40.

[21] Steyerberg EW, Gerl A, Fossa SD, Sleijfer DT, de Wit R, Kirkels WJ, et al. Validity of predictions of residual retroperitoneal mass histology in nonseminomatous testicular cancer. J Clin Oncol 1998;16: 269–74.

[22] Vergouwe Y, Steyerberg EW, Foster RS, Sleijfer DT, Fossa SD, Gerl A, et al. Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: a model update and multicentre validation with more than 1000 patients. Eur Urol 2007;51(2): 424–32.

[23] Van Calster B, Steyerberg EW, D'Agostino RB, Pencina MJ. Sensitivity and specificity can change in opposite directions when new predictive markers are added to risk models. Med Decis Making 2014; 34:513–22.

[24] Collins GS, Moons KG. Comparing risk prediction models. BMJ 2012;344:e3186.

[25] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.

[26] Demler OV, Pencina MJ, D'Agostino RB, Sr. Misuse of DeLong test to compare AUCs for nested models. Stat Med 2012;31:2577–87.

[27] Pepe M, Janes H. Methods for evaluating prediction performance of biomarkers and tests. In: Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A, editors. Risk assessment and evaluation of predictions. New York, NY: Springer New York; 2013:107–42.

[28] Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21(20):3940–1.

[29] Martínez-Camblor P, Pérez-Fernández S, Corral N. Efficient nonparametric confidence bands for receiver operating-characteristic curves. Stat Methods Med Res 2016;27:17.

[30] Steyerberg EW, Vedder MM, Leening MJ, Postmus D, D'Agostino RB, Van Calster B, et al. Graphical assessment of incremental value of novel markers in prediction models: from statistical to decision analytical perspectives. Biom J 2015;57(4):556–70.

[31] Althouse AD. Statistical graphics in action: making better sense of the ROC curve. Int J Cardiol 2016;215:9–10.

[32] 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(1):128–38.

[33] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76.

[34] Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc 2009;172(4):729–48.

[35] 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(6): 796–804.

[36] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6): 565–74.

[37] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6.

[38] Leening MJ, Steyerberg EW, Van Calster B, D'Agostino RB, Pencina MJ. Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. Stat Med 2014;33:3415–8.

[39] Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27: 157–72. discussion 207-12.

[40] Yates JF. External correspondence: decompositions of the mean probability score. Organ Behav Hum Perform 1982;30(1):132–56.

[41] Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation 2009;119:2408–16.

[42] Baker SG, Kramer BS. Peirce, youden, and receiver operating characteristic curves. Am Stat 2007;61(4):343–6.

[43] Pepe MS, Janes H, Li CI, Bossuyt PM, Feng Z, Hilden J. Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? Clin Chem 2016;62:737−42.

[44] Perkins NJ, Schisterman EF. The inconsistency of "optimal" cut-points obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006;163:670−5.

[45] Perkins NJ, Schisterman EF. The Youden Index and the optimal cut-point corrected for measurement error. Biom J 2005;47(4): 428−41.

[46] Morrow DA, Cook NR. Determining decision limits for new biomarkers: clinical and statistical considerations. Clin Chem 2011;57: 1−3.

Update

Journal of Clinical Epidemiology

ERRATUM

Erratum to "ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models" [J Clin Epidemiol. 126C(2020):207-16]

Jan Y. Verbakel[a,b], Ewout W. Steyerberg[c], Hajime Uno[d], Bavo De Cock[e], Laure Wynants[e], Gary S. Collins[f,g], Ben Van Calster[c]

[a]KU Leuven, Department of Public Health and Primary Care, Leuven, Belgium
[b]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[c]Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, the Netherlands
[d]Division of Population Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[e]KU Leuven, Department of Development and Regeneration, Leuven, Belgium
[f]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
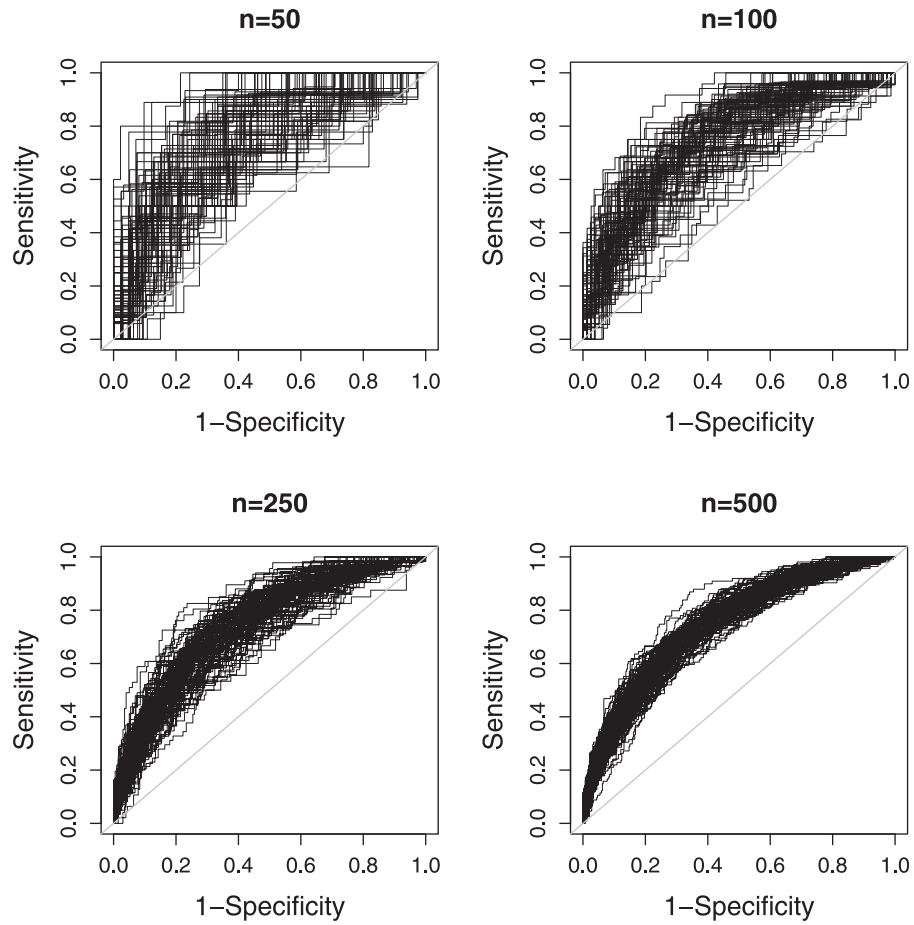[g]Oxford University Hospitals NHS Foundation Trust, Oxford, UK

The publisher regrets that the incorrect figure 1 and figure 4 were published. The article has been corrected.
The publisher would like to apologize for any inconvenience caused.
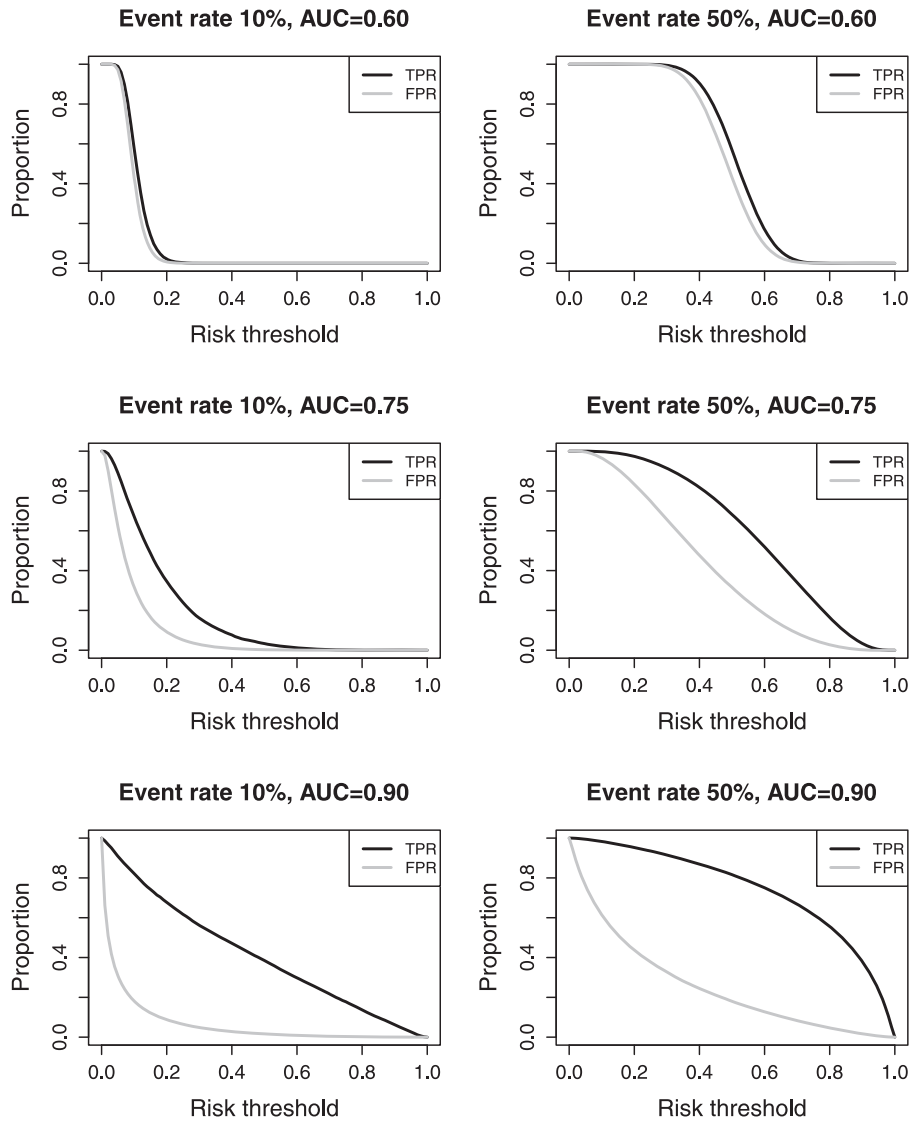
Fig. 1. Theoretical classification plots for scenarios determined by AUCs (0.6, 0.75, or 0.9) and outcome event rate (10% or 50%). The curves assume normally distributed linear predictors.

Fig. 4. Effect of increasing sample size (n) on the variability and shape of the ROC curve. We considered a normally distributed risk prediction model with a true AUC of 0.74 for predicting the presence of an event with a true event rate of 22%. For each of four sample sizes, 100 data sets were simulated and their corresponding ROC curves plotted. At lower sample sizes, ROC curves are more rough and variable than at high sample sizes.