Calibration of risk prediction models: impact on decision-analytic performance[a]

Ben VAN CALSTER[1], PhD, Andrew J. VICKERS[2], PhD

[1] KU Leuven, Department of Development and Regeneration, Leuven, Belgium; [2] Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, USA.

Running head: Calibration and clinical utility of risk models

Word count: 3694

Corresponding author:

Ben Van Calster

KU Leuven, Department of Development and Regeneration

Herestraat 49 box 7003

3000 Leuven

Belgium

T 003216346258

E ben.vancalster@med.kuleuven.be

[a] Financial support for this study provided entirely by a postdoctoral fellowship from the Research Foundation–Flanders (FWO). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. Ben Van Calster is employed by the FWO.

ABSTRACT (270 words)

Decision-analytic measures to assess clinical utility of prediction models and diagnostic tests incorporate the relative clinical consequences of true and false positives without the need for external information such as monetary costs. Net Benefit is a commonly used metric that weights the relative consequences in terms of the risk threshold at which a patient would opt for treatment. Theoretical results demonstrate that clinical utility is affected by a model's calibration, the extent to which estimated risks correspond to observed event rates. We analyzed the effects of different types of miscalibration on Net Benefit, and investigate whether and under what circumstances miscalibration make a model clinically harmful. Clinical harm is defined as a lower Net Benefit compared to classifying all patients as positive or negative by default. We used simulated data to investigate the effect of overestimation, underestimation, overfitting (estimated risks too extreme), and underfitting (estimated risks too close to baseline risk) on Net Benefit for different choices of the risk threshold. In accordance with theory, we observed that miscalibration always reduced Net Benefit. Harm was sometimes observed when models underestimated risk at a threshold below the event rate (as in underestimation and overfitting), or overestimated risk at a threshold above event rate (as in overestimation and overfitting). Underfitting never resulted in a harmful model. The impact of miscalibration decreased with increasing discrimination. Also, Net Benefit was less sensitive to miscalibration for risk thresholds close to the event rate than for other thresholds. We illustrate these findings with examples from the literature and with a case study on testicular cancer diagnosis. Our findings strengthen the importance of obtaining calibrated risk models.

INTRODUCTION

Prediction models that estimate the risk of an event of interest are most commonly evaluated with respect to their capacity to discriminate between events and non-events. This is often evaluated using the area under the receiver operating characteristic curve (AUC), sometimes in combination with citing sensitivity and specificity for one or more thresholds. Calibration is another important aspect of performance, albeit one that has received less attention (1-3). A model is said to be well calibrated if for every 100 patients given a risk of *x%,* close to *x* have the event.

Methodologists have recently criticized discrimination and calibration as being insensitive to clinical consequences (4-7). The consequences of a false negative may be very different from those of a false positive: for instance, diagnosing an invasive tumor as benign, thus delaying appropriate surgery and reducing the chance of cure, is considerably more harmful than an error in the opposite direction, the consequence of which is unnecessary surgery. This has led to the development of decision-analytic measures such as Net Benefit, Relative Utility, and weighted Net Reclassification Improvement (4,6,8). Such measures quantitatively incorporate the relative consequences of false positives and negatives in order to evaluate the potential clinical usefulness of the model. It has been shown that the measures are simple transformations of one another (9), because all use the classic result from decision theory that relative misclassification costs are given by the risk threshold at which one is indifferent about whether or not to take further action such as treatment or further testing (10). As an example, a low risk threshold such as 10% might be used to indicate further testing for cancer, on the grounds that delaying the diagnosis of cancer is far more harmful than unnecessary diagnostic testing; a higher threshold might be used to indicate adjuvant therapy to prevent recurrence, on the grounds that unnecessary chemotherapy is relatively more harmful. Decision-analytic measures are gaining acceptance and popularity (11-14). At the time of

writing, for instance, the paper introducing decision curve analysis (4) has been cited over 200 times. As the decision-analytic measures are interchangeable, we will focus on Net Benefit for simplicity of presentation.

The Net Benefit at a given risk threshold is based on the number of true positives and false positives, where a positive result is defined as a predicted risk greater than the threshold. Following classical decision theory, false positives are then weighted by the odds at the risk threshold. The effects of discrimination on Net Benefit have previously been explored (4,14). Given two well calibrated models, the model with the greater discrimination will have clinical utility across a wider range of risk thresholds, where clinical utility is defined as having a Net Benefit higher than the Net Benefit associated with the default strategies of assuming that all patients are positive ("treat all") or assuming that all patients are negative ("treat none") (Web Figure 1).

It has been shown that Net Benefit is a proper measure of performance (15), which means that Net Benefit obtains its maximum value when the model's risks are correct. Baker and colleagues summarize that utility of prediction using a risk score with threshold T is maximized if the estimated risk at T is accurate, under the assumption that true risk monotonically increases with estimated risk (16). This has important consequences for comparisons between models, for example, when investigating the added value of a new marker over a set of baseline predictors (2,12,17) or when comparing different models for the same disease. However, the relationship between miscalibration and Net Benefit has not been studied in detail.

The aim of the present study is to study how miscalibration affects a model's clinical utility as evaluated by Net Benefit. Of primary interest is the question whether miscalibration can make a model clinically harmful, that is, having a Net Benefit lower than that of either classifying

all patients as positive or all patients as negative. We focus on binary diagnostic or prognostic outcomes predicted by a model that produces a continuous risk score. We perform a simulation study, present a case study on testicular cancer diagnosis, and describe examples from the literature that illustrate our findings.

METHODS

Calibration

Prediction models for outcome Y (1 for event, 0 for non-event) are often defined as well calibrated if the estimated risks of event that they produce are in accordance with the observed conditional event rate. Thus, for patients whose risk estimate R of having the event equals p, $100 \times p$ of 100 should have/develop the event: $P(Y=1|R=p) = p$. A parametric approach to investigate calibration for a specific prediction model using a given dataset is the logistic recalibration model (18). The outcome Y is regressed on the linear predictor (LP) of the model, where the linear predictor is logit(R): $\text{logit}[P(Y)] = a + b \times LP$. Models that are perfectly calibrated for the population from which the patients in the dataset are a representative sample, would have a=0 and b=1. Miscalibration of predicted risks is a common phenomenon when validating models on new populations, for example because of case-mix differences or differences in the effects of the model's predictor variables (19). Generally, four main types of miscalibration can be discerned. If the slope b is lower than 1, the risks are overfitted: small risks are underestimated whereas large risks are overestimated. When b>1, the opposite holds (underfitting). If the intercept deviates from 0 but b=1, the risks are on average overestimated (if a<0) or underestimated (if a>0). More generally overestimation or underestimation is present if $a(b=1)$, i.e. the intercept when the slope is fixed at 1, is different from 0. This can be investigated by adding LP as an offset to the logistic regression model. In simulation studies,

where we control the distribution of Y and of the model predictors, miscalibration can be induced by artificially varying $a(b=1)$ and b.

Net Benefit and decision curve analysis

The choice of risk threshold implicitly conveys the adopted relative misclassification costs. It can be derived that the odds of the risk threshold equals the 'harm-to-benefit ratio', which is the harm of a false positive divided by the benefit of a true positive (9,10). For example, if a risk threshold of 20% is used, the odds are 1 to 4. Therefore, a 20% risk threshold assumes that the harm of a false positive is one quarter of the benefit of a true positive or that one true positive is worth 4 false positives: a clinician might express this in terms such as "I would not do more than five biopsies to find one cancer". Hence, when applying a model to a set of patients, we can correct the number of true positives (TP) for the number of false positive (FP) using the odds of the risk threshold (w): $TP - FP \times w$. When dividing by the total sample size N, the Net Benefit is obtained. The Net Benefit of a model equals the 'net' proportion of true positives, and can be compared to the default strategies of 'treat none' (assume no one has/develops the event) and 'treat all' (assume everyone has/develops the event). Decision-analytic measures such as Net Benefit do not require external information such as monetary costs, hence the definition of Net Benefit in the present context differs from the definition used in the field of cost-effectiveness analysis (20).

By varying the risk threshold, the Net Benefit of the model and the default strategies can be plotted by risk threshold to obtain a decision curve. If, for a given risk threshold, the model has a lower Net Benefit than a default strategy, the model is assumed to be clinically harmful at the implied harm-to-benefit ratio. The Net Benefit of treat none is always 0, whereas Net Benefit of treat all is positive for risk thresholds below the observed event rate in the dataset

and negative for risk thresholds above the observed event rate. When comparing two models, the difference in Net Benefit at a given threshold can be thought of as the difference in the proportion of true positives at the same level of false positives.

Simulation study

In the first part (single model) we considered a model with one predictor. In the second part (model comparison) we considered a baseline model with one predictor and an extended model in which a new marker is added. In the analysis of the single model, the predictor had a standard normal distribution in non-events: $X \sim N(0,1)$. Among events, the predictor was normally distributed with variance of 1 and a mean that was varied to be 0.25, 0.5, 1, 1.5, 2 or 4. These choices correspond to a model with AUCs of 0.57, 0.64, 0.76, 0.86, 0.92, and 0.998. The event rate was varied to be 1%, 20%, or 70%, to approach values seen in many screening, epidemiological, diagnostic, and prognostic studies. In each of the 21 combinations of AUC and event rate we simulated an arbitrarily large sample of 500,000 observations to approach population results. We mathematically derived true risks based on the event rate (ER) and the density of the marker distributions for events and non-events at the marker value as observed for that patient (D_e and D_{ne}, respectively): $ER \times D_e / (ER \times D_e + (1-ER) \times D_{ne})$. This yields perfectly calibrated risks, thus in the recalibration framework the intercept is 0 and the slope 1.

Miscalibration was induced for some models by artificially varying the intercept or slope. General overestimation of predicted risks is obtained by setting $a(b=1) < 0$ and $b=1$. To generate underestimated risks, we set $a(b=1) > 0$ and $b=1$. The $a(b=1)$ values considered are -2, -1, -0.5, 0, 0.5, 1, and 2. To induce overfitting we set $b<1$ while forcing $a(b=1)$ to equal zero; underfitting is achieved by setting $b>1$. The b values considered are 0.25, 0.5, 0.75, 1, 1.5, and 2. Illustrative calibration plots are shown in Web Figures 2 (over- and

underestimation) and 3 (over- and underfitting). These figures are based on an event rate of 20%, and on a model with an AUC of 0.76. Setting $a(b=1)$ to 0 and b to 1 essentially means that we do not change calibration and keep the original, perfectly calibrated risks. Note that the miscalibration induced by either changing the intercept or slope does not affect AUC as the rank order of risks is not changed.

Results were presented as decision curves by varying the choice of risk threshold. For any given setting, a decision curve based on miscalibrated risks was compared with the decision curves based on perfectly calibrated risks, and with decision curves based on the treat all and treat none default strategies.

Sometimes a specific form of miscalibration is encountered in applications where lower predicted risks are calibrated but higher predicted risks are overestimated. This may arise because there is insufficient data on patients with increased risks to accurately estimate these risks. We briefly address this form of miscalibration, focusing on a model that has an AUC of 0.76 to predict an outcome with event rate 1%. Estimated risks above 1% are overestimated.

For the second part of the simulation study we analyze the value of a marker that is added to a baseline model. The baseline and added markers have standard normal distributions in non-events. In events, both are normally distributed with means varied to be 1 (baseline) vs 0.5 (added), 0.5 vs 0.5, or 1 vs 1. AUCs improve by 0.025 (0.760 to 0.785), 0.053 (0.638 to 0.691), and 0.081 (0.760 to 0.841), respectively. The event rate was varied to be 1%, 20%, or 70%. The baseline model is perfectly calibrated, whereas the calibration of the extended model with the added marker is varied. An additional marker can lead to miscalibration for example if the calibration of the marker itself varies between different settings, such as if two laboratories would provide measures 20% apart from the same patient, or if the range of the marker changes between settings. More specifically, apart from perfect calibration $a(b=1)$ is

varied to be -1 and 1, and *b* is varied to be 0.5 and 2. Results are shown as differences in decision curves, i.e. $NB_{extended} - NB_{baseline}$ by risk threshold.

RESULTS

Single model

The results regarding the effect of miscalibration were analogous for all variations of AUC and event rate. Hence we focus on results for 20% event rate, AUC 0.76, $a(b=1)$ values of -1 and -0.5 for overestimation, $a(b=1)$ values of 1 and 0.5 for underestimation, *b* values of 0.25 and 0.5 for overfitting, and a *b* value of 2 for underfitting (Figures 1-2). Similar figures for AUCs of 0.64 and 0.86 are given in Web Figures 4-5.

The overall observation was that miscalibration decreased clinical utility. In general, the impact of miscalibration on utility was lower when discrimination was higher (Web Figures 4-5). Also, Net Benefit was less sensitive to miscalibration for risk thresholds close to the event rate than for other risk thresholds. For models that are overfit or underfit, the calibration curve crosses the ideal line for a risk threshold in the vicinity of the event rate. At this specific point the risk is calibrated, such that the Net Benefit of the model equals that of the perfectly calibrated model. For $b=0.25$ and an event rate of 20%, for example, this occurred at an estimated risk of 28% (Web Figure 3A).

In certain situations we even observed that miscalibration caused the model to be clinically harmful. For models that overestimated risk, i.e. $a(b=1) < 0$, Net Benefit dropped below the treat none performance for a range of thresholds above event rate (Figure 1A). Analogously, for models with underestimated risks Net Benefit dropped below the performance of treat all for a range of thresholds below event rate (Figure 1B). Overfitting of estimated risks

essentially involves underestimation of small risks and overestimation of large risks (Web Figure 3A). Therefore, overfit models yielded Net Benefit values below treat all for a range of thresholds below event rate and below treat none for a range of thresholds above event rate (Figure 2A). Underfitting caused Net Benefit to approach treat all for low risk thresholds and treat none for high risk thresholds (Figure 2B), but never resulted in a harmful model.

When low predicted risks were calibrated but risks above event rate were overestimated, we observed that Net Benefit for risk thresholds above event rate was reduced (Web Figure 6). As expected, this was analogous to results for general overestimation and overfitting that also suffer from the overestimation of high predicted risks.

Model comparison

Given the similarity between simulation scenarios, we focus on results for 20% event rate where the AUC increases from 0.76 to 0.79 when the marker is added (Figure 3), and show other results in Web Figure 7. Results for model comparison nicely fit in with results for single models. Miscalibration of the extended model reduced the difference in Net Benefit with the baseline model. If the extended model underestimates risks, the difference could even become negative (favoring the baseline model) for a range of risk thresholds below event rate. Likewise, if the extended model overestimates risks a negative difference could be observed for a range of risk thresholds above event rate. An overfitting or underfitting extended model lead to a negative difference for a range of thresholds below and above event rate, with the effects being more pronounced for overfitting.

CASY STUDY: TESTICULAR CANCER DIAGNOSIS

After chemotherapy for metastatic nonseminomatous testicular cancer, surgical resection of retroperitoneal lymph nodes is a common treatment because lymph nodes may still contain cancer cells or mature teratoma (21). However, if the lymph nodes merely contain benign tissue there is no clinical benefit from surgical resection. We develop a risk model to predict benign tissue based on data from the Norwegian Radium Hospital (n=150) and validate the model on data from Indiana University Medical Center (n=315). Benign tissue was present in 52% of cases from the development data and in 27% of cases from the validation data. The data were obtained from www.clinicalpredictionmodels.org (14). The five predictors were the maximal diameter of the residual mass, percent reduction in mass size after chemotherapy, presence of teratoma elements in the primary tumor, elevated levels of alpha-fetoprotein, and elevated levels of human chorionic gonadotrophin.

The AUC on the development data was 0.85. The external validation suggests overfitting of the original prediction model, with an AUC of 0.77 and calibration slope *b* of 0.53. A nonparametric calibration curve using loess is shown in Figure 4A. The decision curve, smoothed using a cubic spline, suggests clinical harm at specific thresholds (Figure 4B). For risk thresholds up to 0.15 (below event rate) the model's Net Benefit is below that of treat all, and for risk thresholds between 0.65 and 0.90 (event rate) the Net Benefit is below that of treat none. When recalibrating the model based on Figure 4A, Net Benefit does not fall below the default strategies any more (Figure 4B).

ILLUSTRATIVE EXAMPLES FROM THE LITERATURE

Several published studies provide an empirical demonstration of our findings concerning the effects of miscalibration on Net Benefit. For instance, Collins and Altman validated prognostic models to predict the 10-year risk of cardiovascular disease on a large cohort of

general practice patients in the United Kingdom (22). The incidence of the outcome was 6.6% in women and 8.7% in men. The results showed that the Framingham risk score as modified by NICE (National Institute for Health and Clinical Excellence) overestimates the risk in women and in men, resulting in a harmful model for risk thresholds of around 20% and higher. A second example involves the validation of models to predict prostate cancer in patients referred for prostate biopsy after an abnormal prostate-specific antigen (PSA) level or digital rectal examination (DRE) test (23). The event rate of prostate cancer was 40.7% in the study's validation cohort. The Sunnybrook model underestimated risk for thresholds below the event rate, leading to a harmful model for a range of these thresholds. In turn, the Prostate Cancer Prevention Trial (PCPT) model overestimated risk for thresholds above the event rate, and was harmful within a range of these thresholds.

DISCUSSION

In contrast with statistical measures for discrimination and calibration, decision-analytic measures for clinical utility such as the Net Benefit statistic incorporate clinical consequences (4,6,8,9,24,25). In the present paper we showed the link between calibration and Net Benefit, and in line with theoretical arguments observed that miscalibration always resulted in reduced utility (15,16).

The impact of miscalibration depended on the level and type of miscalibration, the level of discrimination, and the adopted risk threshold (and hence the assumed clinical consequences or harm-to-benefit ratio). We also observed situations where miscalibration of the estimated risks resulted in a clinically harmful model because Net Benefit dropped below one of the default strategies of classifying all patients as positive or as negative. There were two specific situations in which this undesirable result was sometimes observed: (1) when models

underestimated risk at a threshold below the event rate (as in underestimation and overfitting), and (2) when models overestimated risk at a threshold above event rate (as in overestimation and overfitting). In these situations, a clinical harm was less often observed when discrimination performance increased and when the risk threshold was closer to the event rate. An underfit model had lower Net Benefit compared to a well-calibrated model but was never harmful. Thus, when we observe that Net Benefit is below treat all for risk thresholds below the event rate, underestimation of risk can be inferred at these thresholds. Likewise, Net Benefit values below treat none for thresholds above rate are diagnostic of overestimation of risk at these thresholds.

These findings have clinical relevance. Take the case of a risk averse individual who wants to be sure that he has a very low risk of cancer before forgoing further diagnostic work-up. Such an individual would naturally be harmed by being given a risk that is too low, as it may change decision-making inappropriately to no further investigation. On the other hand, a patient who has a high risk threshold is more likely to be harmed by overprediction of risk, as it may lead to interventions that he is disproportionately averse to. Although Net Benefit is less sensitive to miscalibration when the adopted risk threshold is close to the event rate, we still observe decreased clinical utility of decision making at such thresholds under miscalibration.

Analogous observations are made when comparing a baseline model with an extended model that includes an additional predictive marker: the increased utility of the extended model is compromised if the model is miscalibrated, and depending on the situation the extended model may even show less utility than the baseline model. If it would happen that a baseline model is miscalibrated but an extended model is not, opposite results are seen (not shown): the baseline model has decreased utility, resulting in inflated utility of the added marker. This

may occur as a result of poor statistical practice, for example when an already existing baseline model is merely applied but the extended model is fitted on new data.

In the simulation study, we have induced miscalibration through the calibration intercept or slope within a logistic recalibration framework. The only exception is the specific type of miscalibration that is shown in Web Figure 6. We acknowledge that there are other forms of miscalibration that not always fit the parametric framework. Firstly, our conclusions remain for miscalibration on the level of the intercept and slope simultaneously (results not shown). Secondly, it is reasonable to assume that our conclusions also remain for more flexible (non-parametric) forms of the calibration curve (26). Thirdly, a more strict definition of a calibrated risk model requires calibrated risks by covariate pattern rather than simply by the estimated risk (27). Different covariate patterns may result in the same estimated risk, yet this risk may be accurate for some but not all covariate patterns. Further research is needed to elucidate the relationship between calibration and utility when adopting this definition.

In conclusion, miscalibration lowers clinical usefulness and can under specific circumstances yield a model that is clinically harmful at the adopted risk threshold and associated harm-to-benefit ratio. These findings highlight the practical importance of obtaining calibrated risk models.

REFERENCES

1. Collins GS, Moons KGM. Comparing risk prediction models. BMJ. 2012;344:e3186.

2. Tzoulaki I, Liberopoulos G, Ioannidis JPA. Assessment of claims of improved prediction beyond the Framingham risk score. JAMA. 2009;302:2345-2352.

3. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PloS Med. 2012;9:e1001221.

4. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-574.

5. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond'. Stat Med. 2008;27:199-206.

6. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc A. 2009;172:729-748.

7. Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Semin Oncol. 2010;76:1298-1301.

8. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11-21.

9. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. Med Decis Making. 2013;33:490-501.

10. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. New Engl J Med. 1975;293:229-234.

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

12. Moons KGM, de Groot JAH, Linnet K, Reitsma JB, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. Clin Chem. 2012;58:1408-1417.

13. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med. 2012;157:294-295.

14. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer; 2009.

15. Pepe M, Fang J, Feng Z, Gerds T, Hilden J. The net reclassification index (NRI): a misleading measure of prediction improvement with miscalibrated or overfit models. UW Biostatistics Working Paper Series [Internet]. 2013 March [cited 2014 January 2]; Working Paper 392. Available from http://biostats.bepress.com/uwbiostat/paper392.

16. Baker SG, Van Calster B, Steyerberg EW. Evaluating a new marker for risk prediction using the test tradeoff: an update. Int J Biostat. 2012;8:article 5.

17. 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;119:2408-2416.

18. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45:562-565.

19. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971-980.

20. Eckermann S, Briggs A, Willan AR. Health technology assessment in the cost-disutility plane. Med Decis Making. 2008;28:172-181.

21. Vergouwe Y, Steyerberg EW, Foster RS, Sleijfer DT, Fosså 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;51:424-432.

22. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ. 2012;344:e4181.

23. Nam RK, Kattan MW, Chin JL, Trachtenberg J, Singal R, Rendon R, et al. Prospective multi-institutional study evaluating the performance of prostate cancer risk calculators. J Clin Oncol. 2011;29:2959-2964.

24. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest. 2012;42:216-228.

25. Van Calster B, Steyerberg EW, D'Agostino RB Sr, Pencina MJ. Sensitivity and specificity can change in opposite directions when new predictive markers are added to risk models. Med Decis Making. 2014;34:513-522.

26. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33:517-535.

27. Vach W. Calibration of clinical prediction rules does not just assess bias. J Clin Epidemiol. 2013;66:1296-1301.

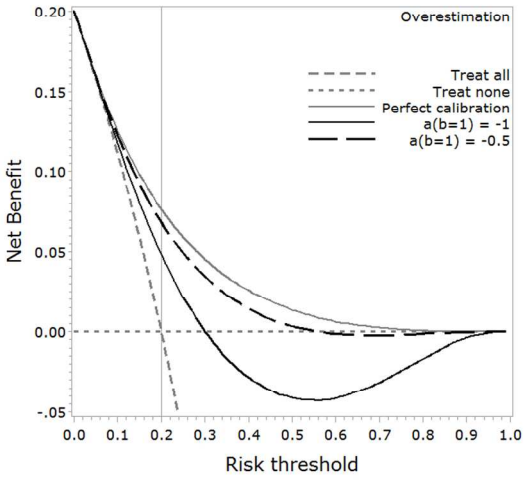
FIGURE LEGENDS

Figure 1. Decision curves for simulated models that generally overestimate (A) or underestimate (B) risk. Event rate is 20%, AUC is 0.76. Results for a perfectly calibrated model are shown as a reference (solid gray lines).

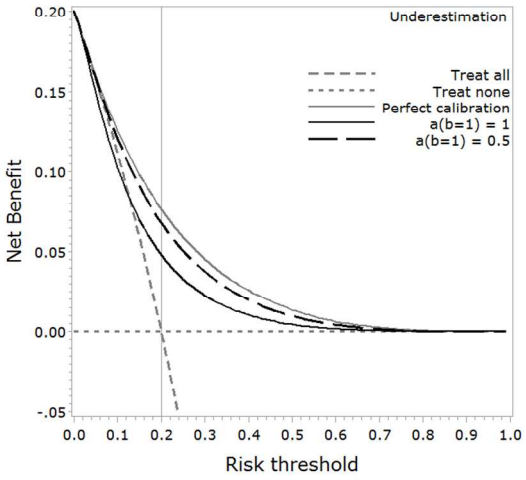
Figure 2. Decision curves for simulated models that overfit (A) or underfit (B) risk. Event rate is 20%, AUC is 0.76. Results for a perfectly calibrated model are shown as a reference (solid gray lines).

Figure 3. Difference in decision curves between a calibrated baseline (AUC 0.76) model and a miscalibrated extended model that includes a novel marker as predictor (AUC 0.79). The effect of four types of miscalibration is shown in comparison to the obtained result had the extended model been perfectly calibrated. The y-axis shows the increase in Net Benefit (NB), $NB_{extended} - NB_{baseline}$, for different choices of risk threshold. Event rate is 20%.

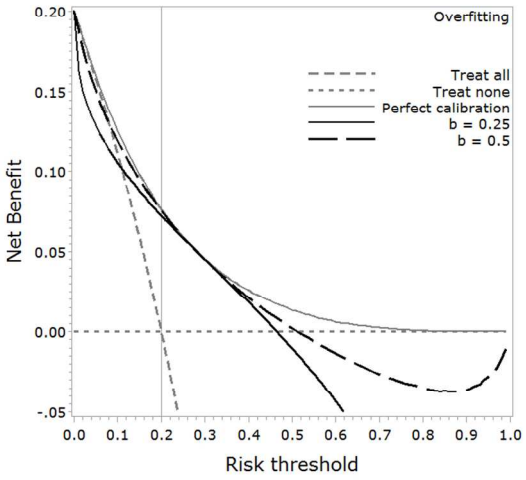
Figure 4. Calibration plot (A) and decision curves (B) obtained on the validation data for the case study on the diagnosis of residual metastatic testicular cancer.



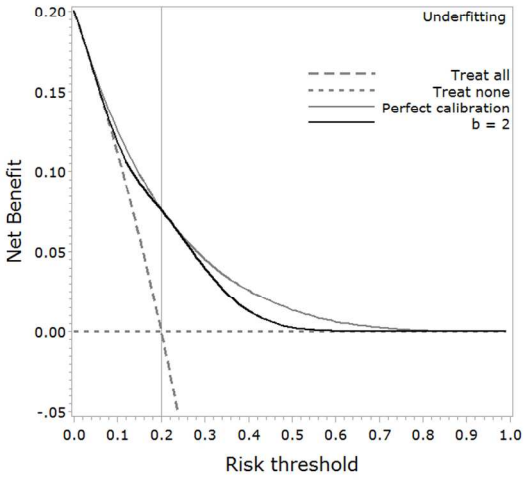
323x181mm (100 x 100 DPI)



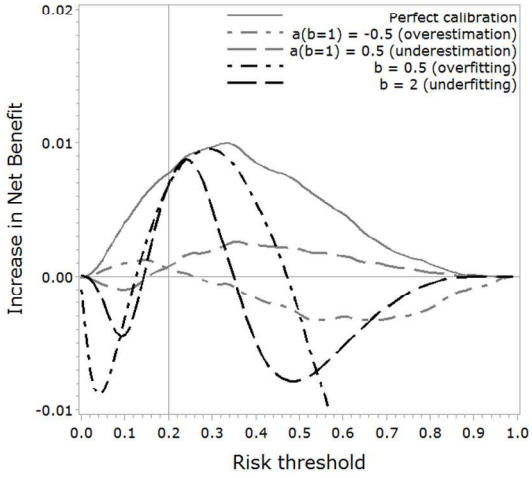
323x181mm (100 x 100 DPI)



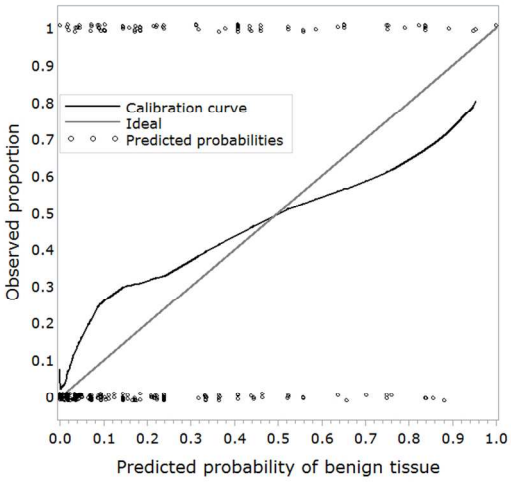
323x181mm (100 x 100 DPI)



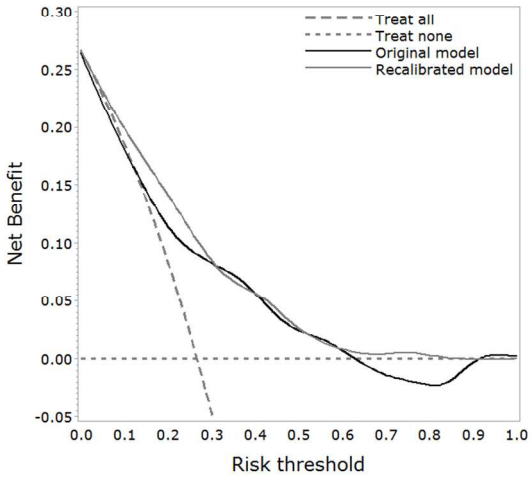
323x181mm (100 x 100 DPI)



323x181mm (100 x 100 DPI)



323x181mm (100 x 100 DPI)



323x181mm (100 x 100 DPI)

Web Figure 1. Decision curves for perfectly calibrated models with varying area under the receiver operating characteristic curve (AUC) when event rate is 20%. The net benefit of the 'treat none' default strategy is 0 irrespective of the adopted risk threshold. The models are based on a single predictor that is normally distributed among events and non-events.

Web Figure 2. Calibration plots for simulated models that generally overestimate ($a(b=1) < 0$) (panel A) or underestimate ($a(b=1) > 0$) (panel B) risk. The event rate is 20%, the AUC 0.76. Results for a perfectly calibrated model are shown as a reference (solid gray lines).

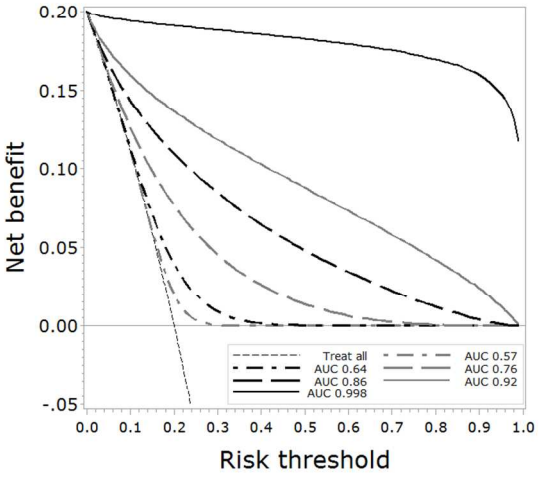
Web Figure 3. Calibration plots for simulated models that overfit ($b < 1$) (panel A) or underfit ($b > 1$) (panel B) risk. The event rate is 20%, the AUC 0.76. Results for a perfectly calibrated model are shown as a reference (solid gray lines).

Web Figure 4. Extra simulation results for a single model. The event rate is 20%, the AUC 0.64. Results for a perfectly calibrated model are shown as a reference (solid gray lines).

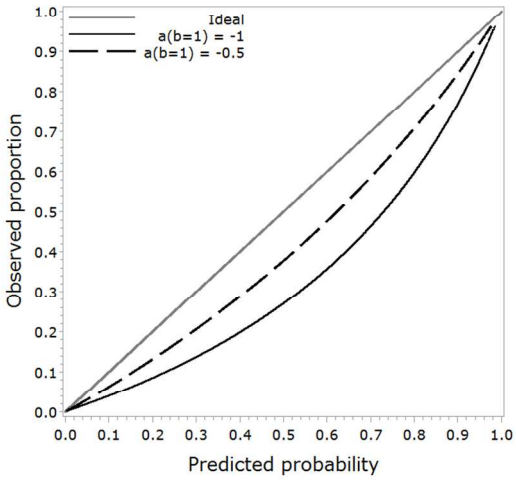
Web Figure 5. Extra simulation results for a single model. The event rate is 20%, the AUC 0.86. Results for a perfectly calibrated model are shown as a reference (solid gray lines).

Web Figure 6. Simulation results for a single model that overestimates high risks. The event rate is 1%, the AUC 0.76.

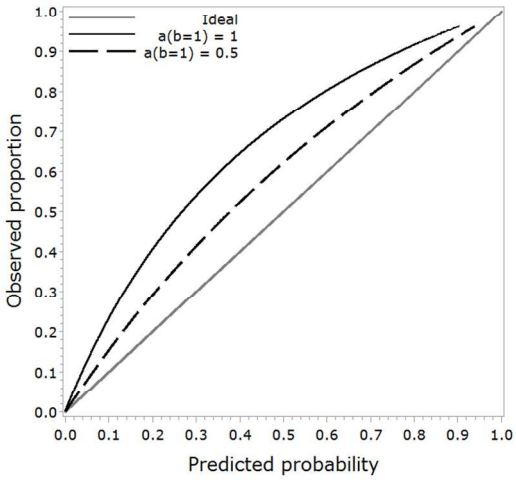
Web Figure 7. Extra simulation results for model comparison. The event rate 20%. The AUC increases from 0.64 (baseline model) to 0.69 (model with novel marker) in panel A, or from 0.76 to 0.84 in panel B.



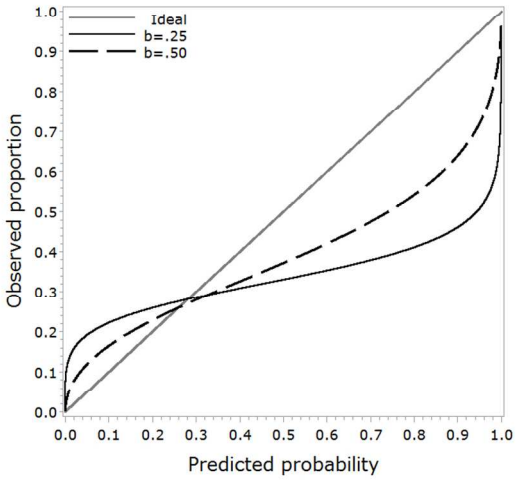
323x181mm (100 x 100 DPI)



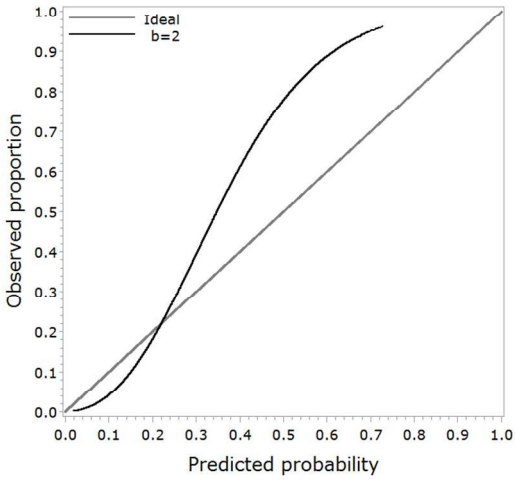
323x181mm (100 x 100 DPI)



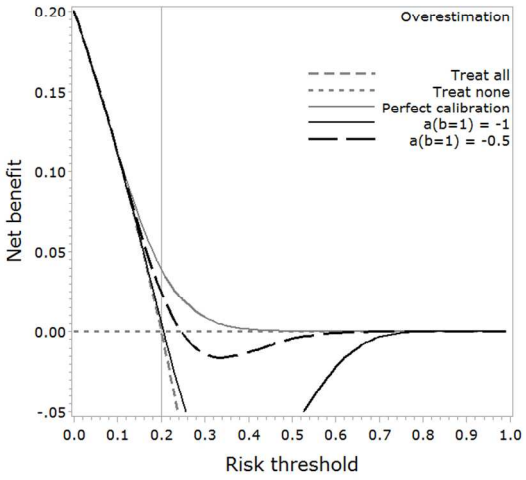
323x181mm (100 x 100 DPI)



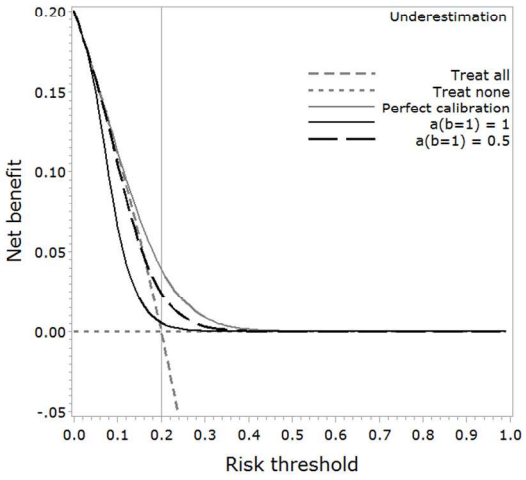
323x181mm (100 x 100 DPI)



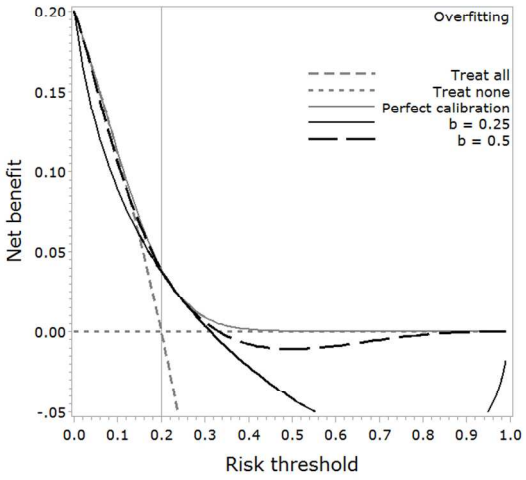
323x181mm (100 x 100 DPI)



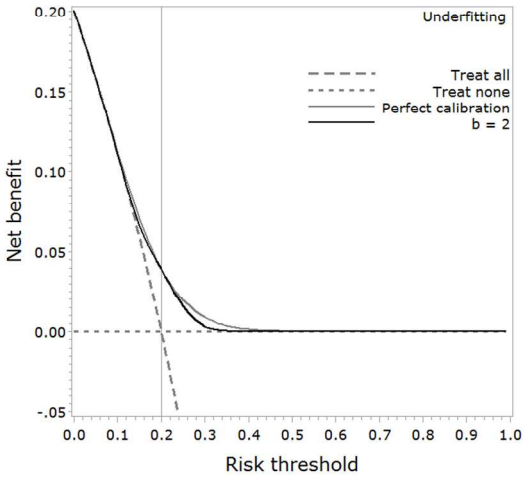
323x181mm (100 x 100 DPI)



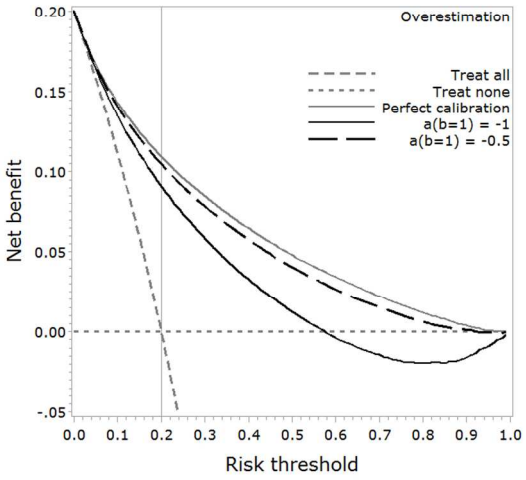
323x181mm (100 x 100 DPI)



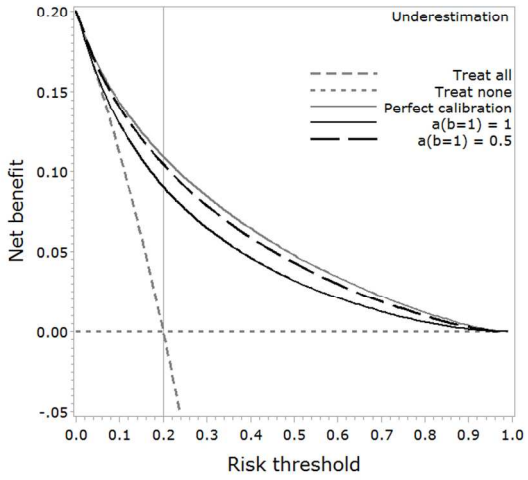
323x181mm (100 x 100 DPI)



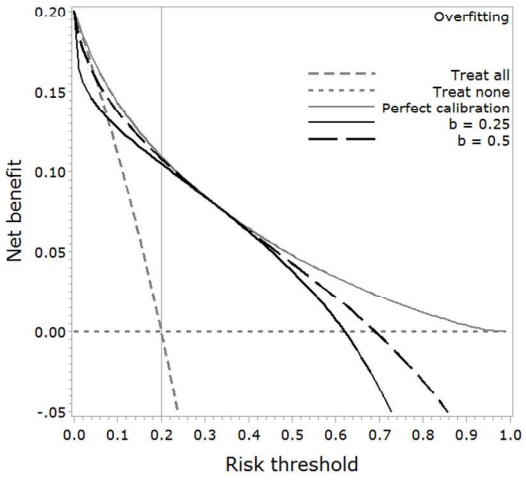
323x181mm (100 x 100 DPI)



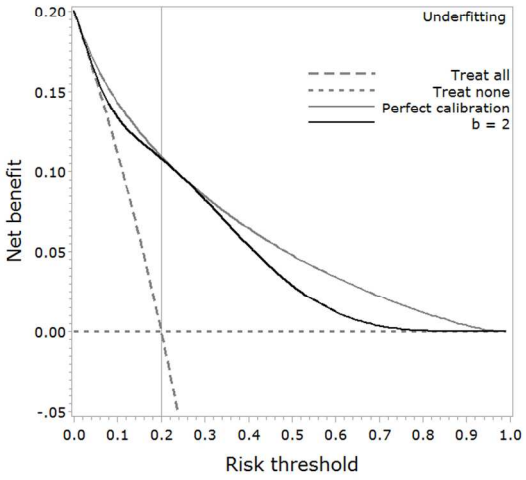
323x181mm (100 x 100 DPI)



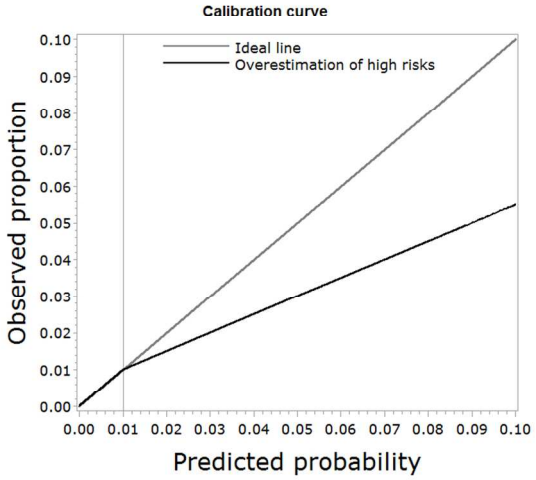
323x181mm (100 x 100 DPI)



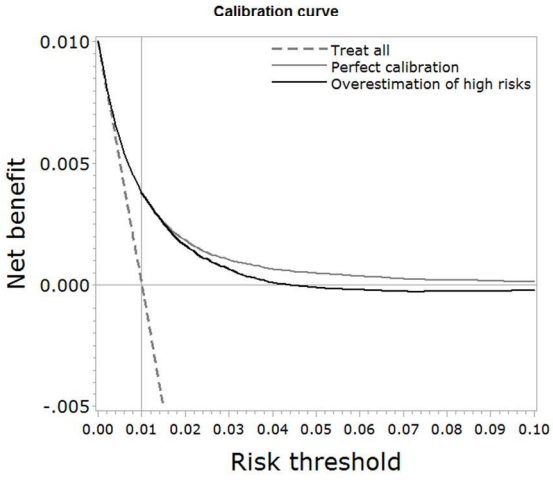
323x181mm (100 x 100 DPI)



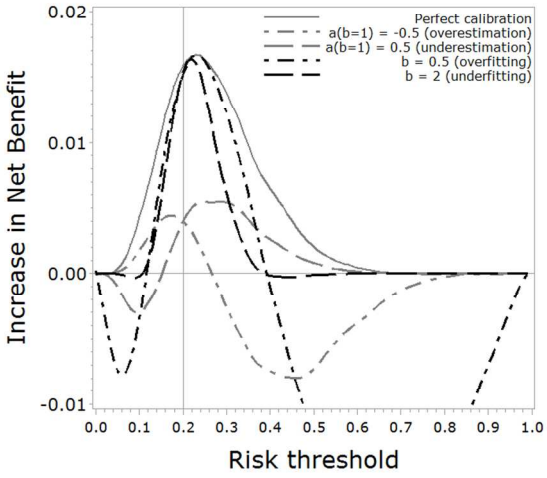
323x181mm (100 x 100 DPI)



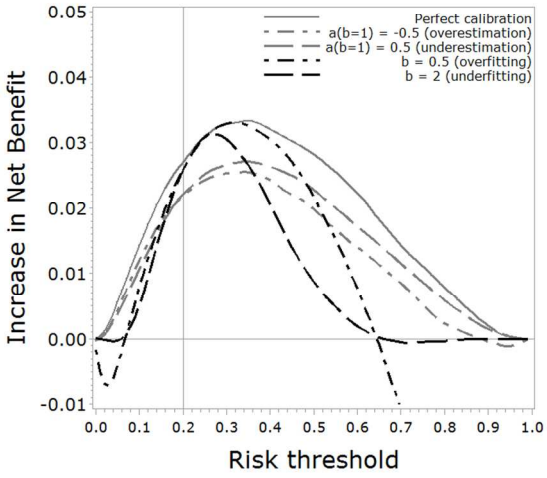
323x181mm (100 x 100 DPI)



323x181mm (100 x 100 DPI)



323x181mm (100 x 100 DPI)



323x181mm (100 x 100 DPI)