Rev Esp Cardiol. 2011;64(9):788–794

Focus on: Contemporary Methods in Biostatistics (IV)

Performance Measures for Prediction Models and Markers: Evaluation of Predictions and Classifications

Ewout W. Steyerberg,[a,*] Ben Van Calster,[a,b] and Michael J. Pencina[c]

[a] Department of Public Health, Erasmus MC, Rotterdam, The Netherlands
[b] Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium
[c] Department of Biostatistics, Boston University; and Harvard Clinical Research Institute, Boston, Massachusetts, United States

Article history:
Available online 16 July 2011

Keywords:
Prediction
Classification
Regression model
Decision analysis

ABSTRACT

Prediction models are becoming more and more important in medicine and cardiology. Nowadays, specific interest focuses on ways in which models can be improved using new prognostic markers. We aim to describe the similarities and differences between performance measures for prediction models. We analyzed data from 3264 subjects to predict 10-year risk of coronary heart disease according to age, systolic blood pressure, diabetes, and smoking. We specifically study the incremental value of adding high-density lipoprotein cholesterol to this model.

We emphasize that we need to separate the evaluation of predictions, where traditional performance measures such as the area under the receiver operating characteristic curve and calibration are useful, from the evaluation of classifications, where various other statistics are now available, including the net reclassification index and net benefit.

Medidas del rendimiento de modelos de predicción y marcadores pronósticos: evaluación de las predicciones y clasificaciones

Palabras clave:
Predicción
Clasificación
Modelo de regresión
Análisis de decisión

RESUMEN

Los modelos de predicción están adquiriendo cada vez mayor importancia en medicina y en cardiología. En la actualidad, hay un interés específico que se centra en las formas de mejorar los modelos con el empleo de nuevos marcadores pronósticos. Nuestro objetivo es describir las semejanzas y diferencias entre las distintas medidas del rendimiento de los modelos de predicción. Hemos analizado los datos de 3.264 individuos para predecir el riesgo de enfermedad coronaria a 10 años, según la edad, la presión arterial sistólica, la diabetes y el tabaquismo. Estudiamos específicamente el valor incremental de la adición a este modelo del colesterol unido a lipoproteínas de alta densidad.

Resaltamos que es preciso separar la evaluación de las predicciones —en las que las medidas de rendimiento tradicionales, como el área bajo la curva *receiver operating characteristic* y la calibración, resultan útiles— de la evaluación de las clasificaciones, para las que disponemos actualmente de otros parámetros estadísticos, como el *net reclassification index* y el beneficio neto.

INTRODUCTION

Prediction models are increasingly important in the medical literature. Many models are available for the prediction of a diagnosis (the presence of disease) and prognosis (for example, incidence of coronary heart disease [CHD]). Quantification of cardiovascular risk is typically accomplished through risk equations or risk score sheets that have been developed from large cohort studies.[1] Modeling techniques include the Cox proportional hazards model and Weibull parametric model.[2]

The Framingham risk functions are among the best known examples of such prediction models.[1,3] They have been essential in individualizing preventive treatment decisions, eg, on using statin therapy. Nowadays, specific interest focuses on ways in which risk prediction can be improved using novel markers[4] identified due to technological advances in basic research, including genomics, proteomics, and noninvasive imaging. These markers hold the promise of bringing personalized medicine closer. An important question is how to evaluate the usefulness of a new marker in making better decisions, such as better targeting of statin therapy to those at increased risk.[5]

A basic condition for a new marker is statistical significance, usually defined as a two-sided P value $<.05$. Statistical significance, however, does not imply clinical relevance, or usefulness of a marker. Indeed, a biomarker with a weak relationship to the outcome of interest can be associated in a statistically significant fashion if examined using a sufficiently large sample size.

* Corresponding author: Department of Public Health, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands.
E-mail address: e.steyerberg@erasmusmc.nl (E.W. Steyerberg).

1885-5857/$ – see front matter
doi:10.1016/j.rec.2011.05.004

Abbreviations

AUC: area under the receiver operating characteristic curve
B: benefit of a true-positive classification
FP: total number of false-positive classifications in the dataset
H: harm of a false-positive classification
NB: net benefit
NRI: net reclassification index
ROC: receiver operating characteristic
TP: total number of true-positive classifications in the dataset

We here aim to describe the similarities and differences between performance measures for prediction models. We are specifically focused on measures to quantify the improvement in predictive performance by adding a marker to an existing prediction model.

METHODS AND RESULTS

Patients

The Framingham Heart Study started in 1948 with a cohort of 5209 individuals. In 1971, 5124 participants (offspring of the original cohort and their spouses) were enrolled in the Framingham Offspring Study. Of these, 3951 participants aged 30 to 74 years attended the fourth cycle of Framingham Offspring cohort examinations, between 1987 and 1992.

As previously described, we excluded participants with prevalent CHD and missing standard risk factors, leaving 3264 of 3951 for the present analysis.[5] Participants were followed for 10 years for the development of coronary heart disease (CHD, including myocardial infarction, angina pectoris, heart failure, or CHD death). A total of 183 subjects developed CHD (5.6%). These data serve as an example to illustrate the concepts rather than to produce a substantive analysis.

Analysis

Cox proportional hazards models were constructed with sex, diabetes, and smoking as dichotomous predictors and age, systolic blood pressure, and total cholesterol as continuous predictors. The hazard ratios were statistically significant for all these predictors. Adding high-density lipoprotein (HDL) cholesterol to this model as a continuous predictor was highly significant (hazard ratio = 0.65, P value $< .001$).[5]

We further focused on the improvement in model performance due to inclusion of HDL cholesterol, comparing 2 sets of predictions of 10-year CHD risk probability: one set of predictions based on a Cox proportional hazards model *without* and one set of predictions based on a model *with* HDL cholesterol included.

Performance Measures for the Quality of Predictions

Discrimination

A key measure for a prediction model is its ability to distinguish those who will develop the event of interest from those who will not; in our case, CHD vs no CHD at 10 years of follow-up.[6] The area under the receiver operating characteristic (ROC) curve (AUC) is the most popular metric to quantify discriminative ability (Table 1).

The ROC curve plots the relationship between sensitivity (or the true-positive rate, ie, the probability of CHD among those classified as positive) and 1 minus the specificity (or the false-positive rate, ie, the probability of no CHD among those classified as negative). The sensitivity and specificity pairs are calculated for all possible cut-offs for the predicted probabilities of 10-year CHD risk. With a low cut-off such as 0.1% risk, the sensitivity is high but specificity is poor. A cut-off of 5.6% corresponds to incidence of CHD (sometimes referred to as "prevalence"). At this cut-off, the model without HDL had a sensitivity of 74% and a specificity of 65% (Fig. 1). The model with HDL performed better at that cut-off (sensitivity 78%, specificity 66%). A higher cut-off such as 20% implied a lower sensitivity but a higher specificity (Fig. 1).

The AUC is equal to the probability that given two subjects (one who developed CHD within the 10-year follow-up and one who did not develop CHD), the model will assign a higher probability of CHD to the former. The AUC for the model without vs that with HDL

Table 1
Some Performance Measures for Prediction Models: Evaluation of Predictions Is Done by Measures Other Than the Evaluation of Better Classification by a Marker

Aspect	Measure	Characteristics
Evaluation of predictions		
Discrimination	AUC or *c* statistic	AUC or *c* is a rank order statistic; Interpretation is as the probability of correct classification for a pair of patients with and without the outcome
Calibration	Intercept and slope of a recalibration model	Intercept ($a|b=1$), reflecting calibration in the large, or the difference between average predictions and average outcome Recalibration slope (b), reflecting the average effect of predictors on the outcome
Evaluation of classifications		
Classification	Youden index	Sum of sensitivity and specificity–1
Clinical usefulness	NB and DCA	Net fraction of true positives gained by making decisions based on predictions at a single threshold (NB) or over a range of thresholds (DCA)
Evaluation of incremental value by a marker		
Increase in discrimination	Delta AUC	Increase in discrimination is usually a modest number
Reclassification	NRI	Net fraction of reclassifications in the right direction by making decisions based on predictions with a marker compared to decisions without the marker
Clinical usefulness	Difference in NB and DCA Weighted NRI	Net fraction of true positives gained by making decisions based on predictions with a marker compared to decisions without the marker at a single threshold (NB) or over a range of thresholds (DCA); weights by consequences of decisions (NB and weighted NRI).

AUC, area under the ROC curve; DCA, decision curve analysis; NB, net benefit; NRI, net reclassification index; ROC, receiver operating characteristic.

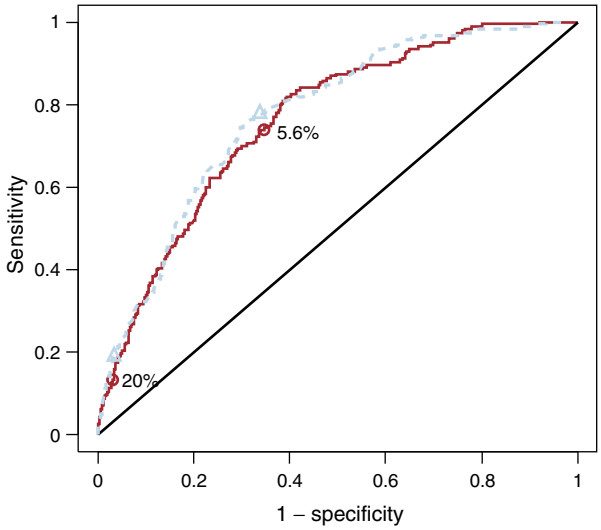

Figure 1. Receiver operating characteristic curves for prediction models of 10 year risk of coronary heart disease based on 3264 subjects. Areas were 0.762 vs 0.774 for the model without vs with high-density lipoproteins. Two cut-offs are shown: 5.6% is the average 10 years incidence of coronary heart disease, and 20% is a clinically relevant cut-off to define high risk subjects.

was 0.762 (95% confidence interval [CI] 0.730–0.794) vs 0.774 (0.742–0.806). This difference of 0.012 is hard to interpret, but would be considered small by most researchers.

Calibration

Another important dimension for the quality of predictions is calibration, ie, agreement between predicted probabilities and observed frequencies of the event of interest.[6] For example, for subjects with a predicted 5% risk of the event of interest, 5 of every 100 subjects, on average, should experience the event. One way to study calibration is to plot a smoothed function of observed events vs predicted probabilities, for example using a loess smoother (Fig. 2).[6] In the ideal case, a 45-degree line is noted, with slope 1 and intercept 0.[2] The slope and intercept can be calculated in a regression model that considers a transformation of the predicted probabilities as the only predictor of the outcome. In our case, we found nearly perfect calibration for a logistic model for 10-year CHD with the logit of the predicted probabilities from the Cox model (Fig. 2).

Graphical Assessment of the Quality of Predictions

In Figure 2, we also show the distributions of predicted probabilities among those with and without CHD to visualize discrimination.[7] There is considerable overlap between these distributions, illustrating what the AUCs of 0.76 and 0.77 mean. The summary measures for this plot can be abbreviated as *a, b,* and *c*: *a* refers to the intercept, or calibration in the large; *b* to the recalibration slope; and *c* to the AUC.[2]

Determining a Cut-off for Classification

The ROC curve considers all consecutive cut-offs to define a high risk vs a low risk group. There are various ways to determine an optimal cut-off. We discuss a data-driven and a decision-analytic (or "utility-based") approach.

Data-Driven Cut-off

A well-known measure for classification performance is Youden's index, which is defined as sensitivity + specificity – 1.[8] Youden's index is maximized in the upper left corner of the ROC curve. So, we might search for the cut-off that corresponds to this point. Interestingly, the point in the upper left corner corresponds to using the incidence of the outcome as the cut-off for the predicted probability, if the prediction model is well calibrated and the ROC curve is concave.[9] In our case this cut-off is 183/3264 = 5.6% (Fig. 1).

Decision-Analytic Cut-off

Decision analysis takes the clinical context as the starting point. The utility, or relative satisfaction, of the consequence of a true or

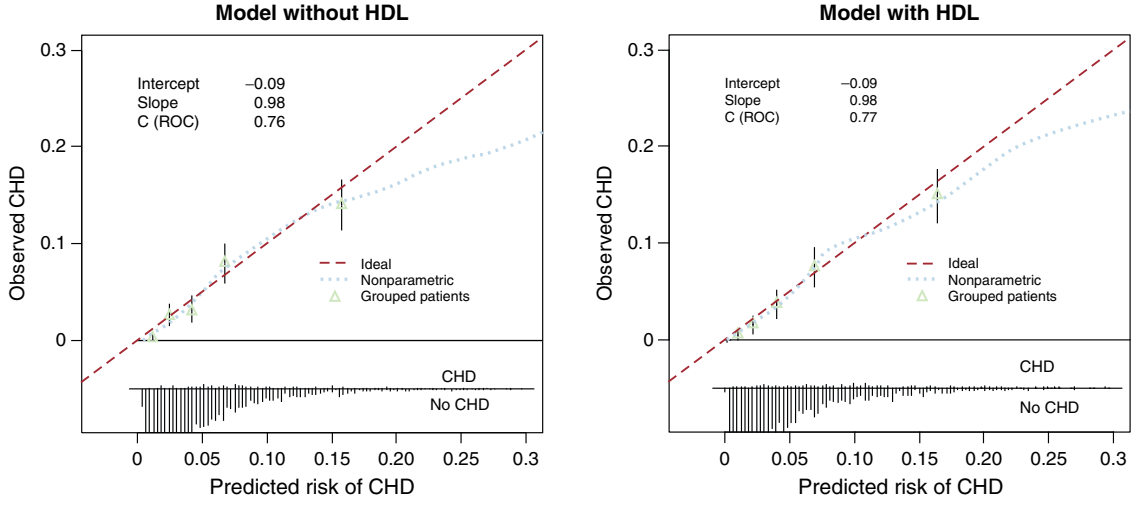

Figure 2. Validation graphs for the model without high-density lipoprotein and with high-density lipoprotein to predict coronary heart disease within 10 years of follow-up. 'Intercept' refers to calibration-in-the-large, and 'slope' refers to the calibration slope for the predictions. 'C (ROC)' refers to the area under the receiver operating characteristic curve. The ideal 45-degree line has intercept 0 and slope 1. Triangles indicate outcomes for quintiles of predictions with 95% confidence intervals. Spikes at the bottom indicate predictions for those with and without coronary heart disease. CHD, coronary heart disease; HDL, high-density lipoprotein; ROC, receiver operating characteristic.

false classification is formally considered.[10] In the case of CHD prevention, a widely accepted cut-off is 20% to define a high-risk group. Formally, this 20% cut-off implies that the utility of false-positive classifications is 4 times less than true-positive classifications, ie, (100 – 20)/20.[7] A false-positive classification implies overtreatment: a subject who will not develop CHD within 10 years is treated, eg, with statins. This harm is weighted as 4 times less important than the benefit of a true-positive classification (a subject who will develop CHD within 10 years is treated with statins). In formula form, the odds of the cut-off equals the harm (H) to benefit (B) ratio:

$$\text{Odds}\,(\text{cut}-\text{off}) = \text{H}/\text{B}.$$

A cut-off of 50% (odds = 1) implies a 1:1 H:B ratio; a 20% cut-off (odds = 1/4) implies a 1:4 ratio. A cut-off of 5.6% maximizes the sum of sensitivity and specificity, but implies that we consider false-positives nearly 20 times less important than true-positives (0.056/0.944).

Performance Measures for the Quality of Classifications

Receiver Operating Characteristic Curves With 1 Cut-off

Rather than considering all possible cut-offs in ROC curves, we can also construct the ROC curves using a single data-driven (Fig. 3A) or decision-analytic cut-off (Fig. 3B). The AUCs are 0.696 and 0.719 for the 5.6% cut-off, and 0.550 and 0.579 for the 20% cut-off, for the model without and with HDL, respectively. Interestingly, the increase in AUC by adding HDL to the prediction model has now increased (from 0.012 for all cut-offs to 0.023 and 0.029 for the 5.6% and 20% cut-offs, respectively).

Reclassification

Cook recognized that a marker's incremental value is expressed in the changes in classifications that occur when predicted probabilities of the marker are considered in the predictive model.[11] For example, considering HDL leads to reclassification of 9.8% of the subjects using the 5.6% cut-off. This number close to 10% is more impressive than the 0.01 increase in AUC over all cut-offs, or the 0.02 increase using the 5.6% cut-off.

Table 2
Reclassification Among 3264 Subjects With and Without a Coronary Heart Disease Event Within 10 Years of Follow-up

		Model without HDL	Model with HDL	
			≤ 5.6%	>5.6%
No CHD (n = 3081)	≤ 5.6%		1872	142[a]
	>5.6%		166[b]	901
CHD (n = 183)	≤ 5.6%		38	10[b]
	>5.6%		3[a]	132

CHD, coronary heart disease; HDL, high-density lipoprotein.
[a] Reclassifications in the wrong direction.
[b] Reclassifications in the right direction.

Net Reclassification

Pencina et al.[5] noted that we should not so much consider reclassification across all patients, but focus on reclassification in the right direction, ie, a higher risk classification for those with CHD and a lower risk for those without CHD. Using the 5.6% cut-off, this net reclassification is 7/183 (3.8%) for those with CHD, and 24/3081 (0.8%) for those without CHD (Table 2). The sum of these numbers is the net reclassification index (NRI): 4.6% [95% CI 0.6%–8.6%]. At the 20% cut-off, NRI = 5.8% [1.4%–10.3%].

Net Benefit

Already in 1884, Peirce[12] stated that the quality of classifications can be expressed as a weighted sum of true-positive classifications: the net benefit (NB). The NB compensates for false-positive classifications by giving these a weight *w*:

$$\text{NB} = (\text{TP} - w\,\text{FP})/\text{N}$$

where TP is the number of true-positive classifications, FP the number of false-positive classifications, and N the total number of subjects.

If $w = 1$, FP and TP are weighted equally. As discussed above, this implies an odds of 1:1 for the H:B ratio. Indeed, *w* is the H:B ratio. Hence, a H:B ratio of 1:4 implies a cut-off of 20% and a 0.25 weight for FP classifications relative to TP classifications, and a 5.6% cut-off implies $w = 0.056/0.944 = 0.059$.

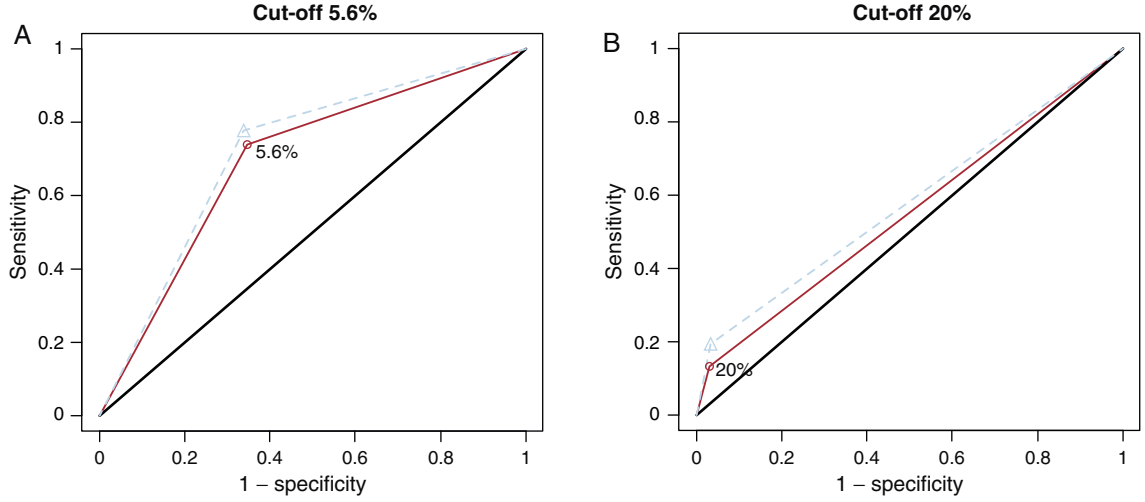

Figure 3. Receiver operating characteristic curves with single cut-offs of 5.6% (A) and 20% (B). The area under the receiver operating characteristic curves are 0.696 and 0.719 for the 5.6% cut-off, and 0.550 and 0.579 for the 20% cut-off, for the model without and with high-density lipoprotein respectively.

Considering the numbers in Table 2, the NB for the model without HDL is calculated as follows: TP = 3 + 132 = 135; FP = 166 + 901 = 1067; w = 0.056/0.944 = 0.059; and N = 3264. This leads to a NB of (135–0.059 × 1067)/3264 = 2.21%. For the model with HDL, the NB is larger: (142–0.059 × 1043)/3264 = 2.47%. The increase in TP is 10–3 = 7, and the decrease in FP classifications is 166–142 = 24. This explains the increase in NB of (7 + 0.059 × 24)/3264 = 0.26%. This number can be interpreted as a net increase in true positive classifications, ie 2.6 more true CHD events are identified per 1000 subjects, at the same number of FP classifications.[13] Equivalently, HDL has to be measured in 1/0.26% = 385 subjects to identify one more TP, using a cut-off of 5.6%.

Decision Curves

The cut-off for clinical application of a prediction model is often not precisely defined. The relative weight of harms and benefits may not be known because of a lack of scientific data, or because of a different appraisal across physicians and patients. Hence Vickers and Elkin[13] proposed to consider a range of cut-offs and calculate the NB across these cut-offs. The result can be plotted in a decision curve (Fig. 4). We note that a small NB is gained by adding HDL to the model for cut-offs between 5% and 25%.

More Cut-offs for Classification

In cardiovascular disease, the use of 3 risk groups is common.[1,5] A low-risk group may be defined as <6% risk, a high-risk group requiring intensive preventive treatment as >20%, with the remainder classified as intermediate risk, requiring lifestyle advice, for example. We can calculate various measures for these 2 cut-offs, including the AUC and NRI. It is not directly possible to calculate NB, since this is defined for 1 cut-off.

We can also consider the whole range of cut-offs for reclassification in a category-less NRI. NRI (>0) is defined as a change in the right direction for any cut-off considered.[14] This calculation should again be considered separately for those with and without CHD. In our case, 62% of the 183 with CHD had higher predictions with the HDL model and 38% had lower predictions, leading to a NRI for events of 24.6%. For the 3081 without CHD, 53% had lower predictions with the HDL model and 47% higher predictions, for a NRI of 5.6%. The NRI (>0) was 0.30. These patterns can also be judged graphically by comparing the predictions with and without HDL in the model in a reclassification plot (Fig. 5)[7,14,15] Here we note that slightly more points fall below the 45-degree line for those with no CHD, and substantially more points fall above the 45-degree line for those with CHD.

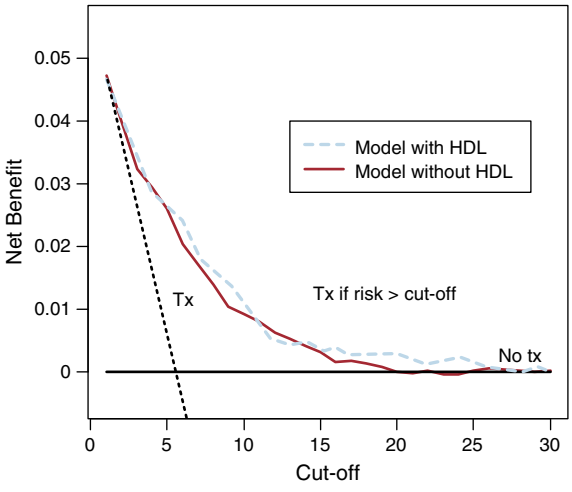

Figure 4. Decision curve for the model without high-density lipoprotein and with high-density lipoprotein to predict coronary heart disease within 10 years of follow-up. The small dotted line indicates the net benefit for "treat all", while the horizontal line indicates "treat none". These 2 lines serve as a reference for the lines for the net benefit of models with or without high-density lipoprotein. HDL, high-density lipoprotein; Tx, treatment.

Interrelationships

If we use a single cut-off, the AUC = (sensitivity + specificity)/2. The increase in AUC (or Δ_{AUC}) is then 0.5 × ($\Delta_{sensitivity}$ + $\Delta_{specificity}$). The NRI in this 2-category case is $\Delta_{sensitivity}$ + $\Delta_{specificity}$, or 2 × Δ_{AUC}.[14] Since Youden index = (sensitivity + specificity)–1, Δ_{Youden} is $\Delta_{sensitivity}$ + $\Delta_{specificity}$; equal to NRI. Indeed the increase in AUC was 0.023 for the 5.6% cut-off, while the NRI and Youden index was 0.046. Hence, it is clear that NRI is a larger number than the increase in AUC.

NRI (>0) is related to Δ_{AUC} over all possible cut-offs. The comparisons used in the calculation of NRI (>0) are between the two prediction models (with and without the marker), but within event groups (CHD, no CHD). Δ_{AUC} is based on pairwise comparisons between event groups (CHD, no CHD) within the two prediction models.[14]

The NB is a weighted sum of sensitivity (fraction TP) and 1–specificity (fraction FP). If the cut-off is the incidence of the outcome, NRI with two categories equals Δ_{NB}/incidence. The 10-year incidence of CHD was 5.6%. Indeed the increase in NB was 0.26% for the 5.6% cut-off, while the NRI was 4.6% (=0.0026/0.056). Hence, it is clear that NRI is a much larger number than the increase in NB. A weighted variant of the NRI has recently been proposed, which behaves similarly to the NB as a summary measure for usefulness of adding a marker to a model.[14]

DISCUSSION

We showed how a number of interrelated measures can be used to indicate the performance of a prediction model. We illustrated the measures with a risk model to predict the 10-year incidence of CHD, with or without using HDL cholesterol as a risk marker. We separated the evaluation of predictions, where traditional performance measures such as the AUC and calibration are useful, from the evaluation of classifications and the contribution of new markers, where various other statistics are now available, including the NRI and NB.[5,7,13,14]

The distinction between a prediction *model* and a prediction *rule* is unclear in most of the current diagnostic and prognostic literature. The key element is that going from a prediction model to a prediction rule requires the definition of a decision threshold, or cut-off.[16] "Prediction model" and "prediction rule" are therefore not synonymous. In a prediction rule, patients with predictions above and below the threshold are classified as positive and negative, respectively. We note that AUC and NRI (>0) evaluate models and not rules. A good model is, however, the first step in creating a good rule.

The threshold for a rule should be appropriate considering the consequences (or *utilities*) of the decision.[10] A false-positive classification (overdiagnosis) is often weighted less in medical contexts than a false-negative classification (underdiagnosis of disease).[16] In the case study, the decision threshold of 20% reflects the 1 to 4 relative weights of false-positive to true-positive classifications. Once the relative weight is used to define the decision threshold, it is logically consistent to also apply this relative weight in the assessment of the quality of decisions. This

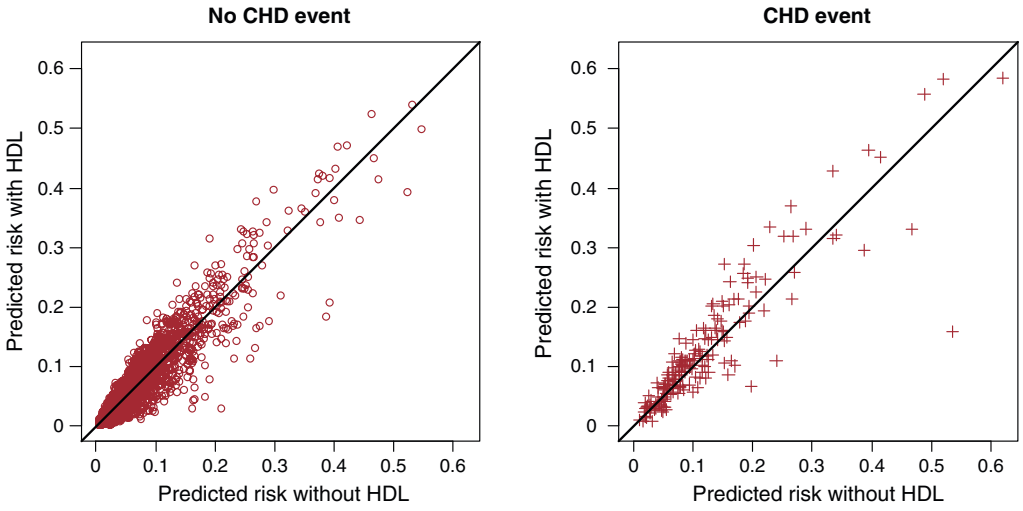

Figure 5. Reclassification plot. CHD, coronary heart disease; HDL, high-density lipoprotein.

principle is followed in the NB definition and in the weighted NRI,[14] and in related measures such as the relative utility.[17] The 2-category NRI is generally not consistent with Δ_{NB} or relative utility. Only if the decision threshold is equal to the incidence of the outcome do NRI and Δ_{NB} give consistent results.

NRI has quickly become popular as a summary measure for the predictive value of a marker. Note that the methodological publications always emphasized the consideration of the separate components of the NRI, ie, NRI for events and NRI for non-events, as shown in Table 2.[5,14]

One reason for the popularity of NRI may be that the absolute number is often given as a percentage, and is then substantially larger than the increase in AUC. In our example, Δ_{AUC} over all cut-offs was 0.012 (Fig. 1), while NRI was +4.6% at a cut-off of 5.6%. Hence NRI is nearly 4 times Δ_{AUC}. However, a fair comparison would consider the cut-off of 5.6% also for Δ_{AUC}, which was 2.3%. Then there is the simple mathematical relationship that NRI = 2 times Δ_{AUC}.[14] Even larger NRI values can be found by considering all cut-offs (NRI [>0] +30%).

Another reason for the popularity of NRI is that AUC is considered "not sensitive" to increases in predictive value of a marker.[11] A recent evaluation found limited statistical power for Δ_{AUC} compared to a likelihood ratio or Wald test for adding a marker to a regression model.[18] These authors however concluded that comparison of AUCs remained useful for initial evaluation of whether a new predictor might be of clinical relevance. There is no reason to assume that the statistical power of NRI is better than a likelihood ratio test; on the contrary, categorizing leads to a loss of predictive information and should lead to less statistical power than a test over the full range of predicted probabilities. In our view, the main issue in performance assessment is not statistical power, but interpretation of the quality of a model and model improvements with markers.

Limitations

Our study has several limitations. We did not use specific methods for survival data, although not all subjects had complete follow-up till 10 years. Censored patients were simply assumed to have no CHD. Methods are available to calculate the AUC (as a concordance, or *c*, statistic) and the NRI for survival data.[14,19] Furthermore, we did not assess the performance as a validation study in independent data. It is common that initial studies of prediction models and markers show promising results, with disappointment in later evaluations. Internal validation with cross-validation or bootstrapping is a minimum requirement.[20] The relatively large sample size (n = 3264 subjects, 183 events) meant that statistical optimism was likely small in our case study (no risk of overfitting), but external validation would be required.

Next to validation and assessment of predictive value, prospective impact studies need to be considered to evaluate the value of prediction models and markers in the improvement of patient outcome.[16] First, we may study whether a model with a marker influences medical decision making compared to a model without the marker. If decision making on further diagnostic work-up or treatment is not different, patient outcomes cannot improve. An ideal study would be a randomized trial on the impact of providing a marker's value on patient outcomes (morbidity, mortality, quality of life), with consideration of process outcomes (diagnostic tests, treatments administered) as intermediate study end points.[4] Since randomized trials may often not be feasible in terms of required research funding and required sample size, formal decision analytic modeling may also be relevant.[21] In such models we can combine estimates of the performance of the prediction model with and without the marker with evidence on the effectiveness of treatment. Treatment could then be more appropriately targeted to those who need it.

CONCLUSIONS

In sum, we recommend the "*a, b, c*" rule for the evaluation of predictions, with *a* (the intercept) and *b* (slope) referring to calibration, and *c* to the AUC (Fig. 2). For the evaluation of classifications and the value of a marker, Δ_{AUC}, event and non-event components of the NRI, NRI (>0), weighted NRI, and NB are appropriate summary measures.

FUNDING

Ewout Steyerberg was supported by the Netherlands Organization for Scientific Research (grant 9120.8004) and the Center for Translational Molecular Medicine (PCMM project). Ben Van Calster has a postdoctoral research grant from the Research Foundation–Flanders (FWO).

CONFLICTS OF INTEREST

None declared.

REFERENCES

1. Pencina MJ, D'Agostino RB, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. Circulation. 2009;119:3078–84.
2. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837–47.
4. 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.
5. 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.
6. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
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.
8. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32–5.
9. Hilden J. The area under the ROC curve and its competitors. Med Decis Making. 1991;11:95–101.
10. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109–17.
11. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115:928–35.
12. Peirce CS. The numerical measure of success of predictions. Science. 1884;4:453–4.
13. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565–74.
14. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11–21.
15. McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med. 2008;168:2304–10.
16. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144:201–9.
17. Baker SG. Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538–42.
18. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Method. 2011;11:13.
19. Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. Ann Intern Med. 2010;152:195–7.
20. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81.
21. Henriksson M, Palmer S, Chen R, Damant J, Fitzpatrick NK, Abrams K, et al. Assessing the cost effectiveness of using prognostic biomarkers with decision models: case study in prioritising patients waiting for coronary artery surgery. BMJ. 2010;340:b5606.