Published in final edited form as:
Med Decis Making. 2013 May ; 33(4): 490–501. doi:10.1177/0272989X12470757.

Evaluation of markers and risk prediction models: Overview of relationships between NRI and decision-analytic measures

Ben Van Calster, PhD, **Andrew J Vickers, PhD**, **Michael J Pencina, PhD**, **Stuart G Baker, ScD**, **Dirk Timmerman, PhD**, and **Ewout W Steyerberg, PhD**
Department of Development and Regeneration, KU Leuven - University of Leuven, Leuven, Belgium (BVC, DT); Department of Public Health, Erasmus MC, Rotterdam, the Netherlands (BVC, EWS); Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, USA (AJV); Department of Biostatistics, Boston University, Boston, USA (MJP); Harvard Clinical Research Institute, Boston, USA (MJP); Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, USA (SGB)

Abstract

BACKGROUND—For the evaluation and comparison of markers and risk prediction models, various novel measures have recently been introduced as alternatives to the commonly used difference in the area under the ROC curve (ΔAUC). The Net Reclassification Improvement (NRI) is increasingly popular to compare predictions with one or more risk thresholds, but decision-analytic approaches have also been proposed.

OBJECTIVE—We aimed to identify the mathematical relationships between novel performance measures for the situation that a single risk threshold *T* is used to classify patients as having the outcome or not.

METHODS—We considered the NRI and three utility-based measures that take misclassification costs into account: difference in Net Benefit (ΔNB), difference in Relative Utility (ΔRU), and weighted NRI (wNRI). We illustrate the behavior of these measures in 1938 women suspect of ovarian cancer (prevalence 28%).

RESULTS—The three utility-based measures appear transformations of each other, and hence always lead to consistent conclusions. On the other hand, conclusions may differ when using the standard NRI, depending on the adopted risk threshold *T*, prevalence *P* and the obtained differences in sensitivity and specificity of the two models that are compared. In the case study, adding the CA-125 tumor marker to a baseline set of covariates yielded a negative NRI yet a positive value for the utility-based measures.

CONCLUSIONS—The decision-analytic measures are each appropriate to indicate the clinical usefulness of an added marker or compare prediction models, since these measures each reflect misclassification costs. This is of practical importance as these measures may thus adjust conclusions based on purely statistical measures. A range of risk thresholds should be considered in applying these measures.

Corresponding author: Ben Van Calster KU Leuven – University of Leuven Department of Development and Regeneration Herestraat 49 box 7003 B-3000 Leuven Belgium Tel +32 16 346258 Fax +32 16 343842 ben.vancalster@med.kuleuven.be.

Introduction

Risk prediction models are essential tools in the era of personalized medicine. Such models provide estimates of diagnostic or prognostic outcomes that can support decision-making in screening, diagnosis and therapy choice across all medical fields (1–3). Often competing models exist for the same outcome, for example a model with or without a novel marker. Novel risk markers hold the promise to refine risk classification which allows better targeting of individuals who will benefit from prevention or therapeutic interventions (4). Another typical example of competing models is when different research groups have independently developed a prediction model.

The evaluation of risk prediction models, and the comparison of competing models, has traditionally focused on discrimination. Hereto, the area under the ROC curve (AUC) is widely used (5). This rank-order statistic assesses how well the model distinguishes between patients with and without the outcome of interest based on the estimated risks. More specifically, the AUC estimates the probability that a patient with the outcome is given a higher risk than a patient without the outcome. The incremental value of a new marker would then be assessed by the difference in AUCs (ΔAUC) for a model with the marker as a predictor variable and a model without (6,7).

However, it is widely recognized that more advanced measures than AUC and ΔAUC are needed (6–11). The main issue is that the AUC lacks clinical interpretability (11,12). It is also unclear what value of ΔAUC is clinically important. Alternative measures have been suggested to evaluate the usefulness of a prediction model in practice, namely assisting with clinical decisions regarding treatment. These include the Net Reclassification Improvement (NRI), weighted NRI (wNRI), Net Benefit (NB), and Relative Utility (RU) (13–17).

This paper investigates the mathematical relationships between these measures. We hypothesize that many of the novel measures are closely related when a single decision threshold is considered. We illustrate the measures with a case study on ovarian tumor diagnosis.

Methods

Terminology and notation

We focus on prediction models for dichotomous outcomes. Extensions to survival type data are possible (15,17,18), but beyond the scope of the present paper. The *prediction model* estimates the risk that the event is present in a given patient. A *prediction rule* is derived from the prediction model through a risk threshold T to classify patients as positive (presence of event) or negative (19). Performance evaluation can focus on predictions or on classifications (6,19,20). We assume a dataset of size N, consisting of N_+ events and N_- non-events. The event prevalence (N_+/ N) is denoted as P. When using threshold T, we denote the number of true positives, false positives, true negatives, and false negatives by **TP**, **FP**, **TN**, and **FN**, respectively. The costs each type of classification is denoted by c_{TP}, c_{FP} c_{FN}, and c_{TN}.

If we have two prediction models, model 1 and model 2, subscripts are used to differentiate between results for the two models (e.g., $\mathbf{TP_1}$ and $\mathbf{TP_2}$). Given the dependence on *T*, the complete notation would be $\mathbf{TP_T}$, $\mathbf{TP}_{\mathbf{1},T}$, etcetera. However, for reasons of simplicity we will omit the conditioning on *T* in the notation.

Classification as a decision problem

For patient classification, we could aim for a threshold that simultaneously optimizes classification of events and non-events (20). However, this strategy does not take misclassification costs into account (21). Disease seriousness, as well as treatment consequences of patient classification, implies that the benefit of accurately classifying diseased patients is often quite different from the harm of assuming disease in non-diseased patients. Usually, benefit of treating a diseased patient exceeds the harm of overtreatment (19). The issue, then, is the choice of a risk threshold that takes the harm-to-benefit ratio into account. The risk threshold can be defined as the risk of the outcome at which one is indifferent about treatment or no treatment, since the expected utility/cost is equal (22). It is well known that there is a direct relationship between the risk threshold and the harm-to-benefit ratio (13,23). If we predict absence of disease in a patient with a risk of disease that equals the threshold *T*, the expected cost will be

$$Tc_{FN}+(1-T)\,c_{TN}. \quad (1)$$

If we predict presence of disease, the expected cost will be

$$Tc_{TP}+(1-T)\,c_{FP}. \quad (2)$$

One is indifferent about treatment if both expected costs are equal. Working out this equality gives

$$T=\frac{c_{TN}-c_{FP}}{(c_{TN}-c_{FP})+(c_{TP}-c_{FN})}. \quad (3)$$

Now, $c_{TN} - c_{FP}$ is the benefit of not treating a non-diseased patient, or equivalently the harm of a false positive. Likewise, $c_{TP} - c_{FN}$ is the benefit if providing treatment to diseased patients, or equivalently the benefit of a true positive. The risk threshold directly informs us about the harm-to-benefit ratio. More specifically,

$$\frac{c_{TN}-c_{FP}}{c_{TP}-c_{FN}}=\frac{T}{1-T}. \quad (4)$$

Thus, the odds of *T* equal the ratio of harm to benefit. This equality was already documented by Pauker and Kassirer (22), and by many researchers thereafter (13,15–17,24–26). For example, adopting a threshold of 0.05 indicates that a true positive is 19 times more important than a false positive, or that a false negative classification is 19 times more costly than a false positive classification. When a threshold of 0.5 is used, both types of misclassifications are considered equally harmful.

Classification based on *T* can be done using the estimated risks from a prediction model, but these risks need to be calibrated to correspond to the intended harm-to-benefit ratio. Calibration means that predicted risks agree with observed outcomes. However, the mathematical relationships between the measures presented in this paper hold irrespective of calibration.

The preferred value of *T* typically varies between individuals. Therefore, as a sensitivity analysis the performance of prediction rules should be reported for different values of *T* (13,17).

Case study: ovarian tumor diagnosis

It is important to accurately characterize an ovarian tumor as benign or malignant prior to surgery. Appropriate treatment of malignant tumors by specialized gynecological oncologists improves prognosis, whereas misdiagnosis of malignancy incurs unnecessary intervention levels. The International Ovarian Tumor Analysis (IOTA) consortium has developed a risk prediction model based on six demographic and ultrasound-based variables to assist clinicians in assessing the risk of malignancy (27). The model does not contain the CA-125 tumor marker, and there is ongoing debate concerning its role in diagnosing ovarian tumors. Therefore, we compare the IOTA model with an extended model that also includes CA-125. The models were developed on data from 1066 patients collected at nine centers between 1999 and 2002, with a 25% prevalence of malignancy. We validated the models on 1938 patients collected at 19 centers between 2005 and 2007, with a prevalence of 28% (28). For each patient the estimated risk of malignancy was obtained from the linear predictor, which was computed by multiplying the estimated regression coefficients with the patients' covariate values. Figure 1(a) shows ROC curves for the models.

False negatives are extremely harmful because patients would not receive appropriate surgery (e.g. laparotomy, interval-debulking) which proved to strongly determine survival of patients. False positives should be avoided in order to save costs and unnecessary examinations and extensive surgery, but they are less harmful. The risk threshold (*T*) adopted by most clinicians and patients will be 10% or lower. We illustrate calculations for *T*=5%, with interest in thresholds between 2 and 10%. In the absence of risk prediction, the default strategy would be to consider all patients as having ovarian cancer and treat them accordingly.

Performance measures

The evaluation of model predictions involves a summary of performance over all possible thresholds. To this end AUC and ΔAUC are traditional measures. When evaluating model classifications, the Youden index is common (29). Recently, NRI has been suggested to compare classifications obtained by two models (14). For a single risk threshold *T*, NRI equals the difference in the Youden index (ΔYouden) (14,15). The Youden index assigns the same importance to the performance for events and non-events and does not account for differential consequences of classifications. It has been labeled a measure of the success or “science of the method” by Peirce (30). In contrast, decision-analytic measures have been proposed that do consider these differential consequences, for example NB (and the difference ΔNB when comparing two models), RU (ΔRU) and wNRI. In Peirce's terminology, these measures consider the “utility of the method” (Table 1).

Net Reclassification Improvement

NRI quantifies the improvement in the reclassification of cases into risk groups when two competing models are compared using the same risk threshold(s) (14). Events should be reclassified to higher risk groups (`up' in equation 5 below), non-events should move in the opposite direction (`down'). For events and non-events separately, the proportion of cases reclassified in the correct direction minus the proportion of cases reclassified in the incorrect direction is computed to obtain the event NRI and non-event NRI. The NRI can then be written as:

$$\begin{aligned}\mathrm{NRI} &= \mathrm{event\ \ NRI\ +\ non{-}event\ \ NRI}\\ &= P\left(\mathrm{up}|\mathrm{event}\right)-P\left(\mathrm{down}|\mathrm{event}\right)+P\left(\mathrm{down}|\mathrm{non{-}event}\right)-P\left(\mathrm{up}|\mathrm{non{-}event}\right).\end{aligned} \quad (5)$$

The NRI can be used for two risk groups based on one risk threshold, but also for three or more risk groups based on two or more thresholds. We use the notation NRI_T as we focus on the NRI for two risk groups using T (15). We can reformulate NRI_T as follows

$$\mathrm{NRI}_T=\frac{\mathbf{TP}_2-\mathbf{TP}_1}{N_+}+\frac{\mathbf{FP}_1-\mathbf{FP}_2}{N_-}=\frac{\Delta\mathbf{TP}}{N_+}+\frac{\Delta\mathbf{FP}}{N_-}. \quad (6)$$

NRI_T is the sum of improvement in sensitivity (event NRI_T) and specificity (non-event NRI_T). For appropriate interpretability, event NRI_T and non-event NRI_T should be reported as well (14,15).

By dividing the difference in true positives by N_+ and the difference in false positives by N_-, NRI_T eliminates the influence of prevalence such that sensitivity and specificity are considered equally important. NRI_T does not account for harms, hence it has the following ambiguity: **TP** and **FP** are a consequence of the chosen threshold T which reflects the harm-to-benefit ratio, but this ratio is not explicitly used to quantify performance. A weighted version of NRI_T has therefore been proposed (15), as discussed later.

As NRI_T equals ΔYouden, Figure 1(b) shows the Youden index of the ovarian tumor models by varying T.

Net Benefit

The NB, written as NB_T to be consistent in notation, corrects **TP** for **FP** based on misclassification costs (13,31), and is written as

$$NB_T=\frac{\mathbf{TP}}{N}-w\frac{\mathbf{FP}}{N}. \quad (7)$$

The benefit of a true positive is considered equivalent to the harm of w false positives. The weight w for false positives is defined as odds(T), or the harm-to-benefit ratio that corresponds to threshold T (13). By penalizing **TP** for w**FP**, NB is the net proportion of true positive classifications. NB_T can be plotted as a function of T, yielding a decision curve. Figure 1(c) shows the decision curves for the ovarian tumor models.

A model can be compared with two baseline strategies, `treat none' (i.e. always predict absence of disease) and `treat all' (i.e. always predict disease). $\mathrm{NB}_{\text{treat none}}$ is always zero as there are no true or false positives. For treat all,

$$NB_{\text{treat all}}=\frac{N_+}{N}-w\frac{N_-}{N}=\frac{P-T}{1-T}. \quad (8)$$

NB_T can be interpreted as the equivalent of the increase in the proportion of true positives relative to `treat none', without increase in false positives. Alternatively, the improvement over $\mathrm{NB}_{\text{treat all}}$ when divided by $w$, $(\mathrm{NB}_T - \mathrm{NB}_{\text{treat all}})/w$, is the equivalent of the decrease in the proportion of false positives relative to `treat all', without decrease in true positives. Decision curve analysis is an elegant way of evaluating the clinical consequences of classifications derived from a prediction model without performing a formal and complex decision analysis (13,31).

The difference in NB_T ($\Delta\mathrm{NB}_T$) of model 1 and model 2 is given as follows:

$$\begin{aligned}\Delta NB &= \left(\frac{\mathbf{TP}_2}{N}-w\frac{\mathbf{FP}_2}{N}\right)-\left(\frac{\mathbf{TP}_1}{N}-w\frac{\mathbf{FP}_1}{N}\right)\\ &=\frac{1}{N}\left(\Delta\mathbf{TP}-w\Delta\mathbf{FP}\right).\end{aligned} \quad (9)$$

$\Delta\mathrm{NB}_T$ can be interpreted as being equivalent to the increase in the proportion of true positives without change in false positives when using model 2 instead of model 1. Alternatively, $\Delta\mathrm{NB}_T/w$ is the equivalent of the decrease in the proportion of false positives without decrease in true positives.

Relative Utility

RU_T is the net benefit in excess of `treat all' or `treat none' (whichever is larger) divided by the net benefit of perfect prediction. RU_T focuses on `treat all' if $T < P$ (or equivalently if treatment is given in the absence of prediction), and on `treat none' if $T \geq P$ (or equivalently if no treatment is given in the absence of prediction). RU_T and NB_T are based on the theory of expected utility, and their detailed description within this framework can be found elsewhere (17,32).

RU_T expresses the utility of a model as the proportion of the maximal gain relative to the best baseline strategy at risk threshold T. RU_T can be plotted as a function of T in a relative utility curve, as illustrated in Figure 1(d) for the ovarian tumor models. The difference in RU_T ($\Delta\mathrm{RU}_T$) of model 1 and model 2 is defined in an analogous manner as for NB_T.

Weighted Net Reclassification Improvement

The critique that NRI does not take cost considerations into account (21,33) led to the introduction of the weighted NRI (wNRI_T when focusing on two risk groups based on T) (15). The savings or benefit when an event is reclassified to a higher risk group by model 2 than by model 1 is denoted by s_1. Likewise, s_2 is used to denote the benefit obtained when a non-event is reclassified to a lower risk group by model 2. When using Bayes' rule on the NRI and including s_1 and s_2, wNRI equals

$$\begin{aligned}\text{wNRI}=&s_1\left(P\left(\text{event}|\text{up}\right)P\left(\text{up}\right)-P\left(\text{event}|\text{down}\right)P\left(\text{down}\right)\right)\\&+s_2\left(P\left(\text{non}-\text{event}|\text{down}\right)P\left(\text{down}\right)-P\left(\text{non}-\text{event}|\text{up}\right)P\left(\text{up}\right)\right).\end{aligned} \tag{10}$$

The ratio of s_2 and s_1, for a prediction rule with two risk groups, correspond to a harm-to-benefit ratio, and was defined as odds(T) to be consistent with the adopted threshold (15). If we set $s_1 = 1/P$ and $s_2 = 1/(1-P)$, we would obtain the original NRI$_T$. For wNRI$_T$, s_1 and s_2 are chosen to maintain the harmonic mean of the original weights $1/P$ and $1/(1-P)$ (15), which is 2. By consequence, $s_1 = 1/T$ and $s_2 = 1/(1-T)$.

Overview of relationships between performance measures

An overview of relationships between measures to quantify performance for dichotomous classifications is presented Table 2. More details on the derivation of these relationships are given in the Appendix. The main finding is that the utility-based measures ΔNB$_T$, ΔRU$_T$, and wNRI$_T$ are transformations of one another, and thus lead to consistent conclusions.

After simple rearrangements of terms, NRI$_T$, ΔNB$_T$, ΔRU$_T$, and wNRI$_T$ are measures with very similar set-up:

$$\begin{aligned}\text{NRI}_T&=\frac{\mathbf{\Delta TP}}{N_+}-\frac{\mathbf{\Delta FP}}{N_-},\\ \Delta NB_T&=\frac{\mathbf{\Delta TP}}{N}-w\frac{\mathbf{\Delta FP}}{N},\\ \Delta RU_{T,T\geq P}&=\frac{\mathbf{\Delta TP}}{N_+}-w\frac{\mathbf{\Delta FP}}{N_+},\\ \Delta RU_{T,T<P}&=\frac{\mathbf{\Delta TN}}{N_-}-\frac{1}{w}\frac{\mathbf{\Delta FN}}{N_-},\\ \text{wNRI}_T&=\frac{1}{T}\left(\frac{\mathbf{\Delta TP}}{N}-w\frac{\mathbf{\Delta FP}}{N}\right).\end{aligned} \tag{11}$$

Alternatively, these measures can be expressed as a weighted sum of the differences in sensitivity and specificity (ΔSe and ΔSp),

$$\begin{aligned}\text{NRI}_T&=\Delta Se+\Delta Sp,\\ \Delta NB_T&=P\Delta Se+w\left(1-P\right)\Delta Sp,\\ \Delta RU_{T,T\geq P}&=\Delta Se+\frac{w(1-P)}{P}\Delta Sp,\\ \Delta RU_{T,T<P}&=\frac{P}{w(1-P)}\Delta Se+\Delta Sp,\\ \text{wNRI}_T&=\frac{P}{T}\Delta Se+\frac{1-P}{1-T}\Delta Sp.\end{aligned} \tag{12}$$

The decision-analytic measures account for disease prevalence and the harm-to-benefit ratio and are thus consistent approaches that are simple transformations of one another. They differ in the specific weights used for **ΔTP** and **ΔFP** (or ΔSe and ΔSp). NB$_\mathrm{T}$ uses weights to obtain a result that can be expressed in basic clinical terms relating to the net number of true positives per 1000 patients. RU uses weights depending on whether or not $T < P$, in order to express NB$_T$ relative to the appropriate baseline strategy and to perfect classification. If $T < P$ NB$_T$ is compared with `treat all' (specificity 0%). Equation 11 shows that RU$_{T,T<P}$ corrects the true negative rate for the cost-adjusted number of false negatives divided by $1 - P$, such that the result is expressed on the scale of specificity. Likewise, if $T \geq P$ NB$_T$ is compared with `treat none' (sensitivity 0%) and RU$_{T,T\geq P}$ corrects the true positive rate for

the cost-adjusted number of false positives divided by P to obtain a result on the scale of sensitivity. Finally, wNRI$_T$ differs from NB$_T$ through the choice of s_1 and s_2 based on a desired harmonic mean of 2.

Maximum test harm of the new marker

When evaluating the added value of a marker, ΔNB$_T$, ΔRU$_T$, and wNRI$_T$ inform on the maximum test harm of the new marker. If the test harm of the marker exceeds its added utility, the model without the new marker may still be preferred. Test harm can be quantified through the test trade-off (32) that is computed as 1/ΔNB$_T$. The test trade-off is the minimum required number of patients tested with the new marker (as part of the extended model) per extra true positive, in order for the increase in net benefit corrected for test harm to be greater than zero. The test trade-off is weighted against the harm associated with obtaining the value of the new marker. For example a test trade-off might be acceptable with a test for a new marker that is not invasive but not acceptable with an invasive test. In other words, if the test trade-off is too high, the added utility of the new marker does not compensate the harm associated with obtaining the marker value.

When are NRI$_T$ and decision-analytic measures inconsistent?

NRI$_T$ and decision-analytic measures can have opposite sign only when the classification of one outcome improves and the classification of the other outcome deteriorates, i.e. when Δ**TP** and Δ**FP** have the same sign. Further, ΔNB$_T$ is 0 when Δ**TP** = wΔ**FP** whereas NRI$_T$ is 0 when $\Delta\mathbf{TP}=\frac{P}{1-P}\Delta\mathbf{FP}$. It follows that, when Δ**TP** and Δ**FP** are both positive, ΔNB$_t$ is positive and NRI$_T$ negative when

$$w<\frac{\Delta\mathbf{TP}}{\Delta\mathbf{FP}}<\frac{P}{1-P}, \quad (13)$$

and the reverse is observed when

$$w>\frac{\Delta\mathbf{TP}}{\Delta\mathbf{FP}}>\frac{P}{1-P}. \quad (14)$$

When Δ**TP** and Δ**FP** are both negative the inequality signs should be switched. This indicates that a large difference between (the odds of) T and P enhances the possibility of contradiction.

Next to the issue of opposite signs, it is also important to discuss inconsistency in strength, i.e. when decision-analytic measures will be strongly supportive of model 2 in case NRI$_T$ is modest or vice versa. We assume the situation that we have computed ΔSe and ΔSp as well as NRI$_T$, and that ΔSe and ΔSp are positive. If at this point a decision-analytic evaluation is desired, P and T come into play. Given ΔSe and ΔSp, equation 12 indicates that the effect of P on ΔNB$_T$ depends on the value of w, more specifically whether or not w > ΔSe/ΔSp. The effect of T is clearer: given ΔSe and ΔSp, ΔNB$_T$ is higher when the adopted T for classification was higher. Then, if NRI$_T$ is supportive of model 2, ΔNB$_T$ can be disappointing if w was low and P was either low (if w < ΔSe/ΔSp) or high (if w > ΔSe/ΔSp).

Let us demonstrate with an example. Assume that ΔSe = ΔSp = 0.05 (both increased with 5 percentage points), such that $\mathrm{NRI}_T = 0.1$. If T was set at 5% and P is 10% the associated $\Delta\mathrm{NB}_T$ is 0.007, suggesting 7 additional true positives per 1000 patients at the same level of false positives. However, if T was set at 20% the $\Delta\mathrm{NB}_T$ increases to 0.016. With T at 5% but P at 50% instead of 10%, $\Delta\mathrm{NB}_T$ even raises to 0.026..

Calibration

The level of calibration affects performance of a prediction rule. With overprediction (estimated risks are too high), the true cut-off is smaller than T, leading to more true and false positives than intended. The opposite, underprediction, is typically worse because some patients would not get the intended treatment. When comparing models, miscalibration of one or both models causes the actually used cut-offs to differ such that models are not compared on equal grounds. Calibration performance is affected by the model and the population to which the model is applied. Therefore it can be argued that its possibly negative effects on decision-analytic measures should play their role. On the other hand it can also be useful to recalibrate a model to assess whether it changes conclusions. Nevertheless, calibration has no effect on the mathematical relationships between measures. Once T is specified, it is used by both models to classify patients and leading to fixed quantities $\mathbf{TP}_1$, $\mathbf{TP}_2$, $\mathbf{FP}_1$, and $\mathbf{FP}_2$.

Results for the case study

Given the skewed distribution of CA-125, the marker was $\log_2$-transformed when adding it to the extended model. The odds ratio was 1.47 (95% CI 1.26 to 1.72), indicating that the odds of malignancy increased by 47% per doubling of CA-125. The AUC increased by 0.008 (95% CI 0.004 to 0.013) from 0.934 to 0.942 (Table 3). This difference cannot be interpreted in a meaningful way, but many investigators would consider this a limited improvement. The $\mathrm{NRI}_{0.05}$ equaled −0.010 (95% CI −0.033 to 0.011): the sum of the differences in sensitivity and specificity decreased with 1.0 percentage points. The sensitivity improved with 1.1 percentage points, but the specificity deteriorated with 2.1 percentage points. The $\Delta\mathrm{NB}_{0.05}$ was 0.0023 (95% CI −0.0011 to 0.0064). This difference can be interpreted as equivalent to an additional 2.3 detected cancers per 1000 patients without an increase in unnecessary invasive surgeries, or, equivalently, to a decrease in 44 unnecessary invasive surgeries per 1000 patients at the same level of detected cancers ($\Delta\mathrm{NB}_{0.05}/w = 0.0437$). Contrary to $\mathrm{NRI}_{0.05}$, $\Delta\mathrm{NB}_{0.05}$ suggested that the CA-125 marker has added value, although the confidence intervals reveal substantial uncertainty in the point estimates. $\mathrm{wNRI}_{0.05}$ is proportional to $\Delta\mathrm{NB}_{0.05}$, so a transformation is needed to obtain the same interpretation. The difference in relative utility was 0.061 (−0.029% to 0.165%), indicating 6.1 percentage points increase in the percentage improvement over `treat all'. Or, as $T < P$, this indicates the increase in net specificity obtained by model 2: at the same level of sensitivity model 2's specificity is 6.1 percentage points higher. The test trade-off derived from the decision-analytic measures was 435. This means that per true positive at least 435 patients need to have CA-125 tested to make risk prediction with model 2 worthwhile.

Figure 2 demonstrates that the decision-analytic measures always give concordant recommendations, but that NRI_T may give a different recommendation. These curves show that the CA-125 marker mainly has value when $T \geq 0.60$. This observation explains why summary measures over all possible thresholds point at an advantage of adding the CA-125 marker. However, high risk thresholds are irrational in this example as it would lead to an unacceptable number of women with cancer who would be denied appropriate treatment.

The calibration plots for both models are shown in Figure 3. After calibration using local regression (loess) as in the calibration plots, sensitivity increased with 0.2 percentage points and specificity with 2.4 percentage points. Then, $\text{NRI}_{0.05}$ equaled 0.026, $\Delta\text{NB}_{0.05}$ 0.0014, and $\Delta\text{RU}_{0.05}$ 0.038. $\text{NRI}_{0.05}$ became more supportive of model 2 whereas decision-analytic measures became less supportive, but now they go in the same direction. The test trade-off was 695 tested patients per extra true positive.

There is publicly available software for net benefit and decision curve analysis (www.decisioncurveanalysis.org; www.clinicalpredictionmodels.org). A simple approach for evaluating a new marker based on relative utility curves and test tradeoff (which is easily extended to survival data and the computation of confidence intervals) is based on calculations involving risk stratification tables (16,32).

Discussion

In this paper we described the mathematical relationships between novel threshold-based measures to evaluate and compare markers and prediction models. The main result is that three recently suggested measures that incorporate misclassification costs, ΔNB_T (13), ΔRU_T (17), and wNRI_T (15) are transformations of each other and hence always lead to consistent conclusions. On the other hand, these conclusions may change when shifting to measures that do not weight binary classifications by misclassification costs. such as with NRI_T. To what extent conclusions differ depends on ΔSe and ΔSp, the adopted risk threshold T and prevalence. Further simulations would be needed to elucidate how common such differences are and what magnitude they may reach.

Following decision theory, we argue that ΔNB_T, ΔRU_T, or wNRI_T be used when classifications obtained by competing models are compared. The adopted T conveys information on the assumed relative misclassification costs. Consequently, when classifications are evaluated without correction for misclassification costs, implicitly assumed information is ignored. To be consistent with decision theory it is essential that the correction for misclassification costs uses the harm-to-benefit ratio assumed by T. Such consistency between classifications and their evaluations is present in NB_T, RU_T, and wNRI_T. To account for variability concerning the preferred risk threshold, T needs to be varied as a sensitivity analysis.

With respect to interpretation, ΔNB_T, ΔRU_T and wNRI_T are different. ΔNB_T can be interpreted in basic clinical terms as the change in the proportion of true positives at the same level of false positives, or, if divided by w, as the change in false positives at the same level of true positives. Even though wNRI_T equals $\Delta\text{NB}_T/T$, its interpretation is less straightforward in clinical terms despite its statistical interpretation. RU_T can be considered as a rescaling of the NB_T relative to the best default strategy and to perfect classification, and thus represents the proportion of the possible improvement over the default strategy that is captured by the model. RU_T can be more specifically interpreted as the net specificity (if $T < P$) or net sensitivity (if $T \geq P$), Then, ΔRU_T reflects the change in specificity at the same level of sensitivity or as the change in sensitivity at the same level of specificity. Therefore, from a decision-analytic perspective, ΔRU_T needs to be interpreted together with the prevalence. E.g. ΔRU_T of 0.1 with $T = 0.2$ has a quite different meaning when prevalence is 1% (ΔNB_T=0.0010; 10 more net true positives per 10,000) or 30% (175 more net true positives per 10,000). However it can be useful to understand the separate contributions of both prevalence and relative utility, as when computing the test tradeoff.

An important aspect of prediction models is the extent to which estimated risks are calibrated, i.e. correspond to observed outcomes. For example, 1 in 20 women with an estimated risk of 5% should have the disease. Calibration problems may especially occur in external validation studies, where systematic differences between predicted risks and observed outcomes are often found (2). If the model is not well calibrated, some patients will not be managed as intended. Such miscalibration also affects model performance using decision-analytic measures as w becomes inconsistent with **TP** and **FP**. This may be seen as problematic, but calibration performance is an inherent part of a prediction model. Miscalibration does not invalidate the mathematical relationships between the measures as described in this work.

A limitation of this work is the focus on classification into two groups. Sometimes more risk groups are desired. We may for example classify patients into a low, intermediate, and high risk group, with further testing in the intermediate risk group (32). Alternatively, each risk group may be associated with a different treatment. A well-known example of such three group classification is in the prevention of cardiovascular disease, with different medical management strategies suggested for different risk categories (14). A strength of the NRI is that it works with any number of risk groups. Research is needed on the extension of decision-analytic measures to situations with three or more risk groups, each with different treatments. Furthermore, the empirical behavior of reclassification measures and decision-analytic measures needs further study, as an extension to the relationships discussed here.

In conclusion, application of prediction models in a decision-making context implies use of a specific risk threshold. Then ΔNB_T, ΔRU_T, or $wNRI_T$ are appropriate measures to indicate clinical usefulness. These novel measures are simple transformations of each other, thus leading to identical conclusions. We recommend using the decision-analytic measures for a range of sensible risk thresholds.

Acknowledgments

Financial support for this study was provided in part by Research Foundation – Flanders (FWO) (1251609N, 1251612N, G049312N), Agency for Innovation by Science and Technology (IWT-Vlaanderen) (grant TBM070706-IOTA3), Center for Translational Molecular Medicine (PCMM project), Prostate Cancer Foundation (funds from David H. Koch), the Sidney Kimmel Center for Prostate and Urologic Cancers, National Cancer Institute (P50-CA92629 SPORE grant to Dr. P. T. Scardino), National Institutes of Health (Stuart G Baker). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. BVC is a postdoctoral fellow of the Research Foundation – Flanders (FWO).

Appendix: derivation of relationships

Net Benefit and Relative Utility

If $T < P$, RU_T can be expressed as

$$\begin{aligned} RU_{T,T<P} &= \frac{NB_T - NB_{\text{treat all}}}{NB_{\max} - NB_{\text{treat all}}} \\ &= \frac{\frac{1}{N}(\mathbf{TP} - N_+ + wN_- - w\mathbf{FP})}{w(1-P)} \\ &= \frac{w\mathbf{TN} - \mathbf{FN}}{wN(1-P)} \\ &= \frac{\mathbf{TN}}{N_-} - \frac{1}{w}\frac{\mathbf{FN}}{N_-}. \end{aligned} \quad (14)$$

If $T \geq P$, RU_T can be expressed as

$$\begin{aligned} RU_{T,T\geq P} &= \frac{NB_T - NB_{\text{treat none}}}{NB_{\max} - NB_{\text{treat none}}} \\ &= \frac{NB_T}{P} \\ &= \frac{\mathbf{TP}}{N_+} - w\frac{\mathbf{FP}}{N_+}. \end{aligned} \quad (15)$$

Thus, if $T \geq P$, RU_T equals NB_T divided by P. If $T < P$, the relationship is slightly more complex. These results lead to the following relationships with $\Delta\mathrm{NB}_T$:

$$\Delta RU_T = \begin{cases} \frac{\Delta NB_T}{w(1-P)} \; if\, T<P \\ \frac{\Delta NB_T}{P} \; if\, T \geq P \end{cases} . \quad (16)$$

Net Reclassification Improvement

NRI_T can be expressed as

$$\begin{aligned} \mathrm{NRI}_T &= \frac{\mathbf{TP}_2 - \mathbf{TP}_1}{N_+} + \frac{\mathbf{FP}_1 - \mathbf{FP}_2}{N_-} \\ &= \frac{1}{N}\left(\frac{1}{P}\mathbf{TP}_2 - \frac{1}{P}\mathbf{TP}_1 + \frac{1}{1-P}\mathbf{FP}_1 - \frac{1}{1-P}\mathbf{FP}_2\right) \\ &= \frac{1}{N_+}\left(\mathbf{TP}_2 - \mathbf{TP}_1 + \frac{P}{1-P}\mathbf{FP}_1 - \frac{P}{1-P}\mathbf{FP}_2\right) \\ &= \frac{1}{N_+}\left[(\mathbf{TP}_2 - \mathbf{TP}_1) - \frac{P}{1-P}(\mathbf{FP}_2 - \mathbf{FP}_1)\right]. \end{aligned} \quad (17)$$

This formula is very similar to $\Delta\mathrm{NB}_T$. The main difference is that P instead of T is used to weight the false positives. Thus, if $T = P$ a special situation occurs when the threshold T equals P, because this makes the NRI_T mathematically proportional to $\Delta\mathrm{NB}_T$ and identical to $\Delta\mathrm{RU}_T$,

$$\mathrm{NRI}_T = \frac{\Delta NB_T}{P} = \Delta RU_T. \quad (18)$$

Also, if $T \geq P$, the NRI_T equals the difference in RU_T for both models with a prevalence-based weight (i.e. odds of P, $P/(1 - P)$) rather than threshold-based weight (w) for the difference in false positives $\mathbf{FP}_2 - \mathbf{FP}_1$.

There is a relationship between NRI_T and ROC curves. The area under the prediction rule's single point ROC curve, AUC_T, is the average of sensitivity and specificity. Hence, NRI_T

equals twice the difference of AUC$_T$ for both models, NRI$_T$ = 2*(ΔAUC$_T$) (10). Links between the Youden index and AUC$_T$ have been reported earlier (34).

Weighted Net Reclassification Improvement

Let $n_{\text{event},ij}$ denote the number of events classified as i = 0,1 by model 1 and j = 0,1 by model 2. Further, $n_{\text{non-event},ij}$ is defined analogously. wNRI$_T$,can then be rewritten as:

$$\begin{aligned}
\mathrm{wNRI}_T &= s_1\left(P(\text{event}\mid\text{up})P(\text{up}) - P(\text{event}|\text{down})P(\text{down})\right) + s_2\left(P(\text{non}-\text{event}|\text{down})P(\text{down}) - P(\text{non}-\text{event}|\text{up})P(\text{u}\right.\\
&= s_1\left(P(\text{up}|\text{event})P(\text{event}) - P(\text{down}|\text{event})P(\text{event})\right) + s_2\left(P(\text{down}|\text{non}-\text{event})P(\text{non}-\text{event}) - P(\text{up}|\text{non}-\text{eve}\right.\\
&= s_1 P(\text{event})\left(P(\text{up}|\text{event}) - P(\text{down}|\text{event})\right) + s_2 P(\text{non}-\text{event})\left(P(\text{down}|\text{non}-\text{event}) - P(\text{up}|\text{non}-\text{event})\right)\\
&= s_1\frac{N_+}{N}\frac{\eta_{event,01}-\eta_{event,10}}{N_+} + s_2\frac{N_-}{N}\frac{\eta_{non-event,10}-\eta_{non-event,01}}{N_-}\\
&= s_1\frac{N_+}{N}\frac{(\eta_{event,11}+\eta_{event,01})-(\eta_{event,11}+\eta_{event,10})}{N_+} + s_2 = \frac{N_-}{N}\frac{(\eta_{non-event,00}+\eta_{non-event,10})-(\eta_{non-event,00}+\eta_{non-event,01})}{N_-}\\
&= s_1\frac{N_+}{N}\frac{\mathbf{TP}_2-\mathbf{TP}_1}{N_+} + s_2\frac{N_-}{N}\frac{\mathbf{FP}_1-\mathbf{FP}_2}{N_-}\\
&= \frac{1}{N}\left(s_1\mathbf{TP}_2 - s_2\mathbf{FP}_2\right) - \frac{1}{N}\left(s_1\mathbf{TP}_1 - s_2\mathbf{FP}_1\right)\\
&= s_1\left(\frac{1}{N}\left(\mathbf{TP}_2 - w\mathbf{FP}_2\right) - \frac{1}{N}\left(\mathbf{TP}_1 - w\mathbf{FP}_1\right)\right)\\
&= s_1\Delta NB_T.
\end{aligned}$$

Thus, wNRI$_T$ is a scaled version of ΔNB$_T$ between the two models. Given that s_1=1/T, the relationship between wNRI$_T$ and ΔNB$_T$ is

$$\mathrm{wNRI}_T = \frac{1}{T}\Delta NB_T. \quad (20)$$

Further, if $T \geq P$ it follows from previous relationships that

$$\mathrm{wNRI}_T = \frac{1}{T}\Delta NB_T = \frac{P}{T}\Delta RU_T. \quad (21)$$

Finally, if $T = P$,

$$\mathrm{wNRI}_T = \mathrm{NRI}_T = \frac{1}{P}\Delta NB_T = \Delta RU_T. \quad (22)$$

Similar to ΔNB$_T$ and ΔRU$_T$, wNRI$_T$ can be plotted as a curve by varying the risk threshold T.

References

1. Vickers AJ. Prediction models in cancer care. CA Cancer J Clin. 2011; 61:315–26.
2. Steyerberg, EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer; New York: 2009.
3. Harrell, FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer; New York: 2001.
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. [PubMed: 19364974]
5. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. J Am Med Assoc. 2009; 302:2345–52.

6. 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. [PubMed: 20010215]

7. 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; 42:216–28. [PubMed: 21726217]

8. Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. Arch Intern Med. 2005; 165:2454–6. [PubMed: 16314539]

9. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol. 2008; 167:362–8. [PubMed: 17982157]

10. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011; 11:13. [PubMed: 21276237]

11. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. Semin Oncol. 2010; 37:31–8. [PubMed: 20172362]

12. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. J Natl Cancer Inst. 2008; 100:978–9. [PubMed: 18612128]

13. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006; 26:565–74. [PubMed: 17099194]

14. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, 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. [PubMed: 17569110]

15. 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. [PubMed: 21204120]

16. Baker SG. Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009; 101:1538–42. [PubMed: 19843888]

17. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc A. 2009; 172:729–48.

18. Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. Ann Intern Med. 2010; 152:195–6. [PubMed: 20124243]

19. 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. [PubMed: 16461965]

20. Steyerberg EW, Van Calster B, Pencina MJ. Performance measures for prediction models and markers: evaluations of predictions and classifications. Rev Esp Cardiol. 2011; 64:788–94.

21. Greenland S. The need for reorientation toward cost-effective prediction. Stat Med. 2008; 27:199–206. [PubMed: 17729377]

22. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med. 1975; 293:229–34. [PubMed: 1143303]

23. Peirce CS. The numerical measure of the success of predictions. Science. 1884; 4:453–4.

24. Elkan, C. The foundations of cost-sensitive learning. In: Nebel, B., editor. Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI); Seattle, USA. 4–10 August 2001; San Francisco: Morgan Kaufmann; 2001:973–8

25. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol. 2002; 20:96–107. [PubMed: 12012295]

26. Hand DJ. Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. Stat Med. 2010; 29:1502–10. [PubMed: 20087877]

27. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. J Clin Oncol. 2005; 23:8794–801. [PubMed: 16314639]

28. Timmerman D, Van Calster B, Testa AC, Guerriero S, Fischerova D, Lissoni AA, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. Ultrasound Obstet Gynecol. 2010; 36:226–34. [PubMed: 20455203]

29. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3:32–5. [PubMed: 15405679]

30. Baker SG, Kramer BS. Peirce, Youden, and receiver operating characteristic curves. Am Stat. 2007; 61:343–6.

31. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Dec Making. 2008; 8:53.

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

33. Vickers AJ, Elkin EB, Steyerberg E. Net reclassification improvement and decision theory. Stat Med. 2009; 28:525–6. [PubMed: 17907248]

34. Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's index. Stat Med. 1996; 15:969–86. [PubMed: 8783436]

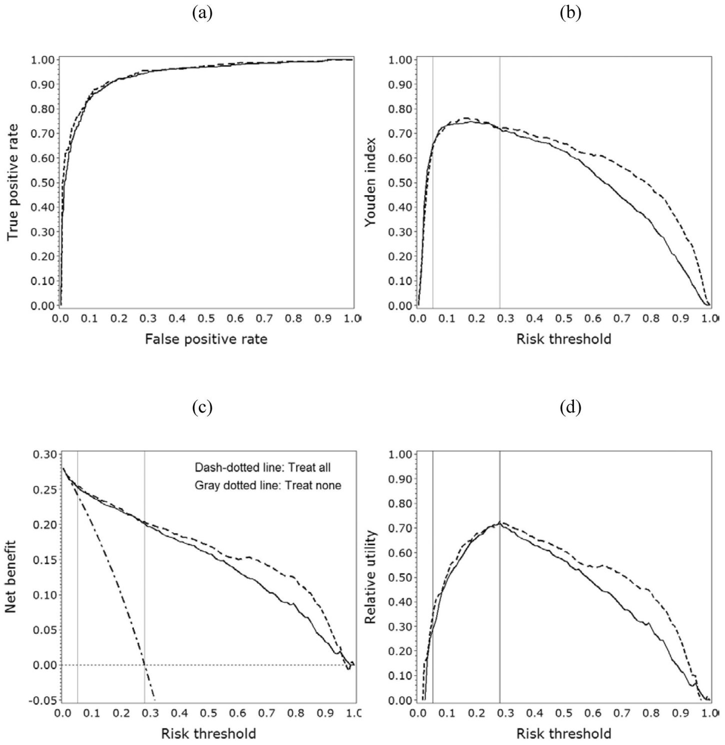

Figure 1.

ROC (panel a), Youden (panel b), decision (panel c), and RU$_T$ (panel d) curves for the reference model (full line) and the extended model (dashed line). Vertical lines in panels b, c, and d indicate the risk threshold *T* of 5% and the prevalence *P* of 28%.

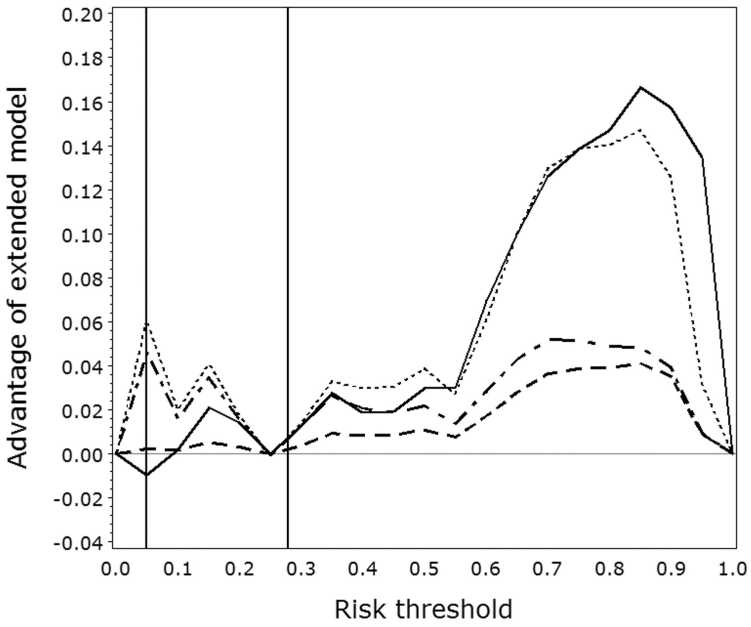

Figure 2.

Curves showing NRI$_T$, ΔNB$_T$, ΔRU$_T$, and wNRI$_T$ (full, dashed, dotted, and dash-dotted lines, respectively) when varying the risk threshold on the x-axis. Vertical lines indicate the risk threshold *T* of 5% and the prevalence *P* of 28%.

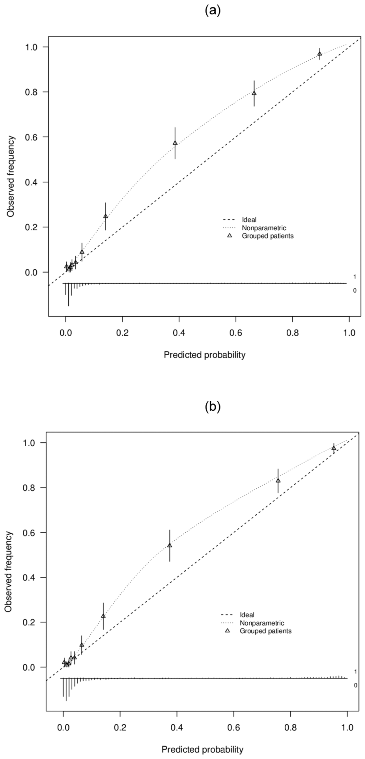

Figure 3.

Calibration plots for the reference model (panel a) and the extended model (panel b).

Table 1

Evaluation measures for the comparison of competing models.

Approach	Method	Characteristics
Evaluation of predictions (summary measures over all possible thresholds)	ΔAUC	Equal to difference in Mann-Whitney statistics; rank order statistic
Evaluation of classifications (using a risk threshold to define two groups)		
`Science of the method'	NRI_T	Sum of differences in sensitivity and specificity; identical to ΔYouden
`Utility of the method'	$\Delta\mathrm{NB}_T$	$(\mathbf{TP} - w\mathbf{FP})/N$
	$\Delta\mathrm{RU}_T$	Expresses NB_T relative to the net benefit of the baseline strategies and the maximum net benefit
	wNRI_T	Weights the NRI_T with misclassification costs

Table 2

Overview of relationships between measures that compare classifications of two competing models at risk threshold *T*.

Condition regarding *T* and *P*	**Relationships between measures**
If $T < P$	$\Delta RU_T = \frac{1}{w(1-P)} \Delta NB_T$
If $T \geq P$	$\Delta RU_T = \frac{1}{P} \Delta NB_T$
	$\text{wNRI}_T = \frac{P}{T} \Delta RU_T$
If $T = P$	$\text{NRI}_T = \frac{1}{P} \Delta NB_T = \Delta RU_T = \text{wNRI}_T$
Irrespective of *T* and *P*	$\text{NRI}_T = \Delta\text{Youden}$
	$\text{NRI}_T = 2^*\Delta\text{AUC}_T$ (i.e. twice the difference in the areas under the prediction rules' single point ROCs)
	$\text{wNRI}_T = \frac{1}{T} \Delta NB_T$

Table 3

Evaluation measures for the reference and extended models for the diagnosis of ovarian tumors. The risk threshold *T* is set at 0.05.

Approach	Method	Value (95% CI)[*]	Reference vs extended model
Evaluation of predictions (summary measure)	ΔAUC	0.008 (0.004; 0.013)	AUC: 0.934 vs 0.942
Evaluation of classifications			
Science of the method	$\text{NRI}_{0.05}$	−0.010 (−0.033; 0.011)	Youden index: 0.654 vs 0.644
	Event $\text{NRI}_{0.05}$	0.011 (0.000; 0.025)	Sensitivity: 0.943 vs 0.954
	Non-event $\text{NRI}_{0.05}$	−0.021 (−0.040; −0.004)	Specificity: 0.711 vs 0.691
Utility of the method	$\Delta\text{NB}_{0.05}$	0.0023 (−0.0011;0.0064)	$\text{NB}_{0.05}$: 0.253 vs 0.255
	$\Delta\text{RU}_{0.05}$	0.061 (−0.029; 0.165)	$\text{RU}_{0.05}$: 0.289 vs 0.350
	$\text{wNRI}_{0.05}$	0.046 (−0.022; 0.127)	

[*] Approximate confidence intervals were obtained using the bias-corrected bootstrap method based on 1000 replicates of the dataset.