DOI: 10.1111/j.1365-2362.2011.02562.x

METHODS

Assessing the incremental value of diagnostic and prognostic markers: a review and illustration

Ewout W. Steyerberg[*], Michael J. Pencina[†], Hester F. Lingsma[*], Michael W. Kattan[‡], Andrew J. Vickers[§] and Ben Van Calster[*,¶]

[*]Department of Public Health, Erasmus MC, Rotterdam, The Netherlands, [†]Department of Biostatistics, Boston University, and Harvard Clinical Research Institute, Boston, MA, USA, [‡]Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA, [§]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA, [¶]Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium

ABSTRACT

Background New markers may improve prediction of diagnostic and prognostic outcomes. We review various measures to quantify the incremental value of markers over standard, readily available characteristics.

Methods Widely used traditional measures include the improvement in model fit or in the area under the receiver operating characteristic (ROC) curve (AUC). New measures include the net reclassification index (NRI) and decision-analytic measures, such as the fraction of true-positive classifications penalized for false-positive classifications [net benefit (NB)]. For illustration, we discuss a case study on the presence of residual tumour vs. benign tissue in 544 patients with testicular cancer. We assessed three tumour markers [Alpha-fetoprotein (AFP), Human chorionic gonadotropin (HCG) and Lactate dehydrogenase (LDH)] for their incremental value over currently standard clinical predictors.

Results AUC and R^2 values suggested adding continuous LDH and AFP whereas NB only favoured HCG as a potentially promising marker at a clinically defendable decision threshold of 20% risk. The NRI suggested reclassification potential of all three markers.

Conclusions The improvement in standard discrimination measures, which focus on finding variables that might be promising across all decision thresholds, may not detect the most informative markers at a specific threshold of particular clinical relevance. When a marker is intended to support decision-making, calculation of the improvement in a decision-analytic measure, such as NB, is preferable over an overall judgment as obtained from the AUC in ROC analysis.

Keywords Incremental value, logistic regression model, performance measures, prediction.

Eur J Clin Invest 2011

Introduction

Novel markers are being identified in large numbers nowadays following technological advances in basic research, including genomics, proteomics and noninvasive imaging. These markers hold the promise of improving the prediction of diagnostic and prognostic outcomes and bring personalized medicine closer [1]. Despite their importance to medical care, methods for evaluation of the performance of markers are still underdeveloped [2].

It has been emphasized before that the incremental value of a marker over standard, readily available diagnostic characteristics is of key interest [3,4]. Ideally, a previously published prediction model is available as a reference model in the analysis. For example, the value of markers for cardiovascular disease may be studied in a statistical model that includes predictors identified in the Framingham study [5]. A recent review, however, found that the exact definition of the reference model varied substantially across studies that claimed to adjust for 'Framingham predictors', with better performance for markers when added to poorer performing reference models [6].

In this paper, we aim to review the properties of a number of traditional and relatively novel measures to evaluate the predictive performance of a marker. We use a case study on markers for patients with testicular cancer to illustrate the behaviour of different performance measures and to highlight some general methodological challenges in assessing the incremental value of a diagnostic marker.

Clinical example

Men with metastatic nonseminomatous testicular cancer can nowadays often be cured by cisplatin-based chemotherapy. After chemotherapy, surgical resection is a generally accepted treatment to remove remnants of the initial metastases, because residual tumour tissue (residual cancer cells or mature teratoma) may still be present. In the absence of tumour tissue, resection has no therapeutic benefits, while it is associated with hospital admission and risks of morbidity and mortality. Currently, resection is usually advised if the postchemotherapy size of a residual tumour mass exceeds 10 mm. More diagnostic characteristics have, however, been described, including the reduction in mass size, the histology of the primary tumour and three tumour markers [Alpha-fetoprotein (AFP), Human chorionic gonadotropin (HCG) and Lactate dehydrogenase (LDH)] [7]. We focus on the incremental value of these three markers in predicting the residual histology of 544 patients, where 299 had residual tumour and 245 benign tissue [8].

All analyses were performed in R version 2.11.1 (R Foundation for Statistical Computing, Vienna, Austria), using the Design library. The syntax and data are publicly available at http://www.clinicalpredictionmodels.org.

Statistical modelling

We consider the situation that we are interested in the value of a test or marker in predicting the presence or outcome of a disease. We aim to determine the incremental value of the marker over other predictors, often including demographics (age and sex) and other basic characteristics (e.g. history, presenting signs and symptoms) [9]. For dichotomous outcomes, multivariable logistic regression analysis is a standard statistical technique to achieve this aim [10]. The basic effect measure from a logistic regression model is the odds ratio (OR). Predictions of the outcome can be calculated based on the odds ratios of the predictors in the model and the model intercept [11].

Several methodological issues arise in such multivariable regression analyses, including the coding of a marker and the choice of the reference model. The specific focus of this paper is on measures of overall predictive performance, improved classification ('discrimination') and improved decision-making ('clinical usefulness', see Table 1).

Coding of continuous markers

Markers measured on an ordinal or continuous scale are often dichotomized, such that we can consider them as 'positive' vs. 'negative'. Although this practice makes interpretation of the effect of a marker straightforward, it implies a loss of

Table 1 Characteristics of some measures to quantify the incremental value of a diagnostic marker

Aspect	Measure	Characteristics
Independent association	Odds ratio (OR)	Quantifies relative risk, either for the marker alone (univariate analysis) or additional to other predictors of outcome (multivariable, or adjusted, analysis). For a binary marker, the OR refers to the comparison of a positive vs. a negative marker value. For a continuous marker, the OR refers to a one unit increase in marker value
Overall performance	Difference in Nagelkerke R^2, Pearson R^2, or Brier score Integrated discrimination improvement (IDI)	Better with lower distance between observed and predicted outcome IDI equals the difference in Pearson R^2
Discrimination	Difference in area under the receiver operating characteristic curve (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
Reclassification	Net reclassification index (NRI)	Net fraction of reclassifications in the right direction by making decisions based on predictions with the marker compared to decisions without the marker; default weights are by prevalence of disease.
Clinical usefulness	Difference in net benefit (NB) and decision curve analysis (DCA) Weighted NRI	Net fraction of true positives gained by making decisions based on predictions with the 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)

information [12]. The alternative of considering continuous versions of a marker poses the challenge of careful handling potential nonlinearity in the relationship between the marker and the disease. One common transformation is to take the logarithm of a marker value, which may especially be useful for skewed distributions. Alternatives include polynomials such as the square root, square or cubic transformations. More flexible functions may also be considered, such as 'fractional polynomials' [13] or spline functions [14]. Especially, restricted cubic spline functions are attractive, because these provide a family of flexible forms without capitalizing on chance findings in the data under study (using few degrees of freedom) [15].

We considered the relationship of the marker LDH to the presence of residual tumour in the case study. We first examined nonlinearity with a flexible spline function (Fig. 1). Lower values of LDH are associated with a higher likelihood of residual tumour at resection. The logarithm of LDH was subsequently used, because a linear effect of the log-transformed LDH reasonably approximated the spline function. A dichotomization of LDH as lower vs. higher than the upper limit of normal was also considered for comparison of how much information is lost by dichotomization.

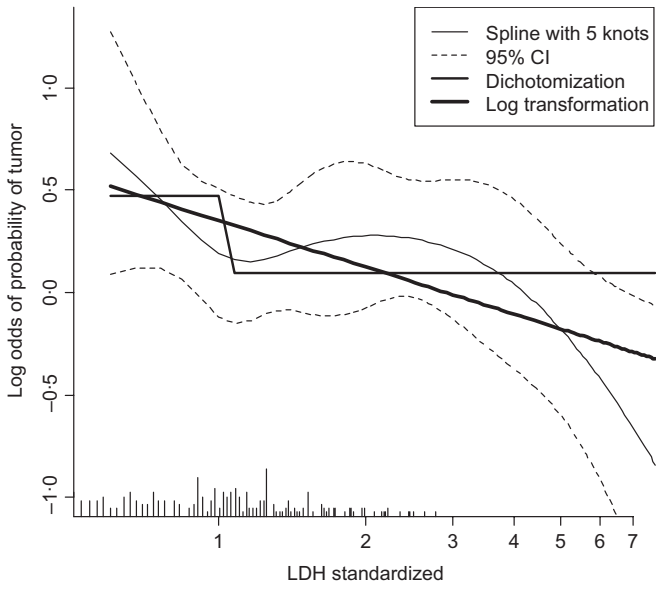

Figure 1 Relationship of LDH to presence of residual tumor at resection in testicular cancer patients. A restricted cubic spline function was used with five knots (shown with 95% confidence intervals), a dichotomized version (LDH elevated vs. normal), and a logarithmic transformation. Note that the *x*-axis is log transformed such that a straight line is shown for log(LDH). The distribution of LDH values is indicated by spikes at the bottom of the graph.

Multivariable analysis and the choice of reference model

A simple first step is to perform a univariate analysis for the marker, i.e. without any further adjustment for patient or disease characteristics. We should, however, be more interested in the incremental value of a marker, on top of predictors that are readily available [3]. It is common to consider additional value over a set of 'established predictors', preferably in a previously published prediction model. In cardiovascular disease, it is common to consider prediction models developed with the Framingham cohort [5], although several other models are available. Other models are common to take as a reference in other fields, e.g. the Gail [16] model in breast cancer research.

We consider two reference models in the case study to illustrate the relevance of using a more extensive reference model rather than a limited one. These reference models are a multivariable combination of postchemotherapy size, reduction in size and primary histology vs. postchemotherapy size alone. The odds ratios of AFP and HCG were between 2 and 3, either in univariate or adjusted analyses, and always highly statistically significant (Table 2). Because AFP and HCG are commonly considered as elevated vs. normal, we did not attempt to model these markers as continuous predictors, in contrast to LDH. The odds ratio for normal LDH was relatively small in univariate analysis (OR = 1·5, P = 0·055) and larger when adjusted for postchemotherapy size (OR = 2·6, P < 0·001) or three other characteristics (OR = 1·9, P = 0·013). A similar pattern was noted for the continuous version of LDH, where odds ratios were calculated for the 25 vs. the 75 percentile to allow for a fair comparison to the dichotomized markers. *P*-values were lower for the continuous version of LDH, reflecting the fuller use of information in the statistical analysis.

Assessing overall incremental value

The distance between the predicted outcome ($\hat{Y}$) and actual outcome (Y) is central to quantify overall model performance from a statistical modeller's perspective [17,18]. For binary outcomes, we define Y as 0 or 1 and $\hat{Y}$ as the predicted probability P. A model with a marker added should have a smaller distance between predicted and observed outcomes.

Explained variation (R^2) can be calculated for generalized linear models [19]. One common option is to use Nagelkerke's R^2 [11,20]. This measure is based on a rescaling of the fit of the model according to the −2 log likelihood. Another option is to simply calculate Pearson R^2. This R^2 measure considers the squared distances between predictions p and the outcome Y. Pearson R^2 is hence related to measures such as the Brier score, which also considers such squared distances [17].

The area under the receiver operating characteristic (ROC) curve (AUC) is the most commonly used performance measure

Table 2 Odds ratios for three tumor markers in logistic regression models in testicular cancer data set (n = 544), without any other predictors (univariate), with statistical adjustment for postchemotherapy size, and with statistical adjustment for postchemothrapy size, reduction in size, and primary tumor histology (presence of teratoma). The outcome was the presence of residual tumor at postchemotherapy resection (299/544, 55%)

Characteristic	Univariate	Adjusted for postchemotherapy size	Adjusted for postchemotherapy size, reduction, and primary histology
Prechemotherapy AFP elevated	2·8 (2·0–4·1)	2·2 (1·5–3·3)	2·7 (1·7–4·2)
Prechemotherapy HCG elevated	2·2 (1·5–3·1)	2·0 (1·3–2·9)	2·1 (1·4–3·2)
Prechemotherapy LDH normal	1·5 (1·0–2·1)	2·6 (1·7–4·1)	1·9 (1·1–3·0)
log(LDH/upper limit of local normal value)*	1·4 (1·1–1·8)	2·9 (2·1–4·1)	2·1 (1·5–3·1)

*Values are odds ratios with 95% confidence intervals for comparison of the 25 to the 75 percentile. LDH was first studied with a restricted cubic spline function, which could be approximated well with a log transformation.

to indicate the discriminative ability of a prediction model. The ROC curve is a plot of the sensitivity (true-positive rate) against 1 – specificity (false-positive rate) for consecutive cut-offs for the probability of the outcome. AUC is identical to the concordance index (c), which is a rank-order statistic for predictions p against actual outcomes Y [11]. The AUC or c can be interpreted as the probability that the patient with a higher predicted probability has the outcome, when we consider a pair of patients of one with and one without the outcome. Useless predictions such as a coin flip result in an AUC of 0·5, while a perfect prediction model has an AUC value of 1.

To assess incremental performance, the difference in R^2 or c statistics is commonly considered, comparing a model with the marker to a model without [6,9,21]. We studied uncertainty in the differences with a bootstrap procedure, where patients were sampled with replacement. Models were refitted in each bootstrap sample to estimate the standard error (SE) of the distribution of each performance measure [18]. We calculated 95% confidence intervals around the original estimates as ±1·96 SE. These intervals do not include zero for statistically significant differences at the 0·05 level.

R^2 and AUC in the case study

The increases in Nagelkerke R^2 values for dichotomized markers were up to 8% in univariate analyses and around 3% in adjusted analyses. As expected, the continuous version of LDH had larger R^2 values than its dichotomized version in all analyses. The best performance was noted for AFP in univariate and fully adjusted analyses, while continuous LDH performed best when adjustment was only for postchemotherapy size (Table 3a).

Receiver operating characteristic curves were constructed for models with and without tumour markers (Fig. 2). Larger improvements in AUC are noted when only postchemotherapy size was modelled as a reference (Fig. 2a) compared to taking the model with the three predictors postchemotherapy size, reduction and primary histology as a reference (Fig. 2b). This illustrates that the reference model is an important issue in judging the incremental value of a diagnostic marker.

The increase in AUC followed the same pattern as for the R^2 values. Increases were between 0·01 and 0·02 for the fully adjusted analyses, where measurement of AFP and continuous LDH contributed most to improving discrimination between those with and without residual tumour (Table 3b).

Reclassification and clinical usefulness

Novel measures related to reclassification

A 'reclassification table' shows how many subjects are reclassified by adding a marker to a model [22]. For example, a model with traditional risk factors for cardiovascular disease was extended with the predictors 'parental history of myocardial infarction' and 'C-reactive protein (CRP)'. The increase in c statistic was minimal (from 0·805 to 0·808). However, when the predicted risks were categorized with three cut-offs into four groups (0–5%, 5–10%, 10–20%, > 20% 10-year cardiovascular disease risk), about 30% of individuals changed category when comparing the extended model with the traditional one. Change in risk categories, however, is insufficient to evaluate improvement in risk stratification; the changes must be appropriate. An 'upward' movement in categories for subjects with the outcome implies improved classification, and any 'downward movement' indicates worse reclassification. The interpretation is opposite for subjects without the outcome. The overall improvement in reclassification can be quantified as the sum of differences in proportions of individuals moving up minus the proportion moving down for those with the outcome, and the proportion of individuals moving down minus the proportion moving up for those without the outcome, which has been referred to as the Net Reclassification Index (NRI) [23].

The NRI was introduced with an example in cardiovascular disease prevention, where three risk categories are commonly considered (0–6%, 6–20%, > 20%) [23]. A category-free version

Table 3 Performance of testicular cancer models with or without the tumor makers AFP. HCG, and LDH according to Nagelkerke's R^2 (a) and *c* statistics (b)

Characteristic	Univariate	Compared to postchemotherapy size	Compared to postchemotherapy size, reduction, and primary histology
(a)			
Reference value	0%	22·9% (15·1–30·6%)	34·1% (26·3–41·9%)
AFP abnormal	7·7% (+7·7%) (4·7–10·7%)	26·0% (+3·1%) (0–6·2%)	37·8% (+3·7%) (0–6·9%)
HCG abnormal	4·7% (+4·7%) (2·0–7·4%)	25·4% (+2·5%) (−0·1–5·2%)	36·3% (+2·2%) (−0·2–4·7%)
LDH abnormal continuous	0·9% (+0·9%) (−1·9–3·9%) 1·5% (+1·5%) (−3·3–6·3%)	26·8% (+3·9%) (0–6·9%) 31·6% (+8·7%) (3·9%–13·5%)	35·2% (+1·1%) (−0·2–2·9%) 37·1% (+3·0%) (−0·1–5·8%)
(b)			
Reference value	0·5	0·748 (0·707–0·790)	0·794 (0·756–0·832)
AFP abnormal	0·616 (+0·116) (0·068–0·164)	0·764 (+0·016) (−0·001–0·033)	0·814 (+0·019) (0·001–0·035)
HCG abnormal	0·592 (+0·092) (0·043–0·140)	0·761 (+0·013) (−0·002–0·027)	0·804 (+0·010) (−0·003–0·021)
LDH abnormal continuous	0·537 (+0·037) (−0·010–0·084) 0·550 (+0·050) (0·002–0·099)	0·769 (+0·021) (0·004–0·039) 0·793 (+0·045) (0·019–0·072)	0·799 (+0·005) (−0·005–0·015) 0·811 (+0·017) (0·002–0·033)

Nagelkerke's R^2: Values are Nagelkerke's R^2 values (partial R^2). Reference values for comparison are in the first row. Values between brackets are 95% confidence intervals.
c statistics: Values are AUC values (improvement in AUC). Reference values for comparison are in the first row. Values between brackets are 95% confidence intervals.

has advantages if categories are less strongly defined, and when comparisons are to be made between studies [24]. The formulas remain the same when using the category-less NRI(> 0), but the definition of upward or downward movement is simplified to indicate any increase or decrease in probabilities of the outcome. Another option is to calculate the integrated discrimination improvement (IDI), which also considers improvements over all possible categorizations. IDI is identical to the difference in Pearson R^2 values and relates to the difference in discrimination slopes of predictions based on models with and without the marker [23,25].

Novel measures related to clinical usefulness

In the calculation of the NRI with two categories (high risk vs. low risk), the improvement in sensitivity [true positives (TP)] and the improvement in specificity (true negatives) are summed. This implies relatively more weight for detecting disease if disease was less common than no disease. For example, even if the prevalence of a disease is 1%, the improvement in TPs is weighted the same as the improvement in true negatives (see Appendix). Hence, weighting is based on the prevalence of disease and not on clinical consequences. It is hence informative to study the individual components of the NRI (one for events and one for non-events) and not only their sum [26].

The net benefit (NB) is a measure that explicitly incorporates weights for detecting disease (TP) vs. overdiagnosing nondisease [false positives (FP)] [27]. NB is defined as: $NB = (TP - wFP)/N$, where N is the total number of patients and w is the relative weight for overdiagnosis (FP) vs. appropriate diagnosis (TP) [27,28]. The NB can be interpreted as the fraction of TP classifications penalized for FP classifications. The NB indicates how many more TP classifications can be made with a model for the same number of FP classifications, compared to not using a model [27].

In the case of a marker, classifications may sometimes be pre-defined as positive vs. negative. But when a marker is added to a reference model, we will usually obtain a risk function with probabilities for the outcome under study. Classification of individuals is then based on a decision threshold on the probability scale, p_t. The additional value of a marker can then be summarized as the difference in NB (ΔNB) at p_t for predictions made with and without using the marker in the risk function.

Interestingly, the threshold p_t by definition reflects the relative weight for false-positive vs. true-positive classifications [28]. Hence, the weight w in the NB formula directly corresponds to the decision threshold p_t. More specifically, w equals $p_t/(1 - p_t)$, implying that w is equal to the odds of the decision threshold p_t. For example, a decision threshold of 20% implies that FPs are valued at 1/4th of detecting disease or another outcome, and $w = 0{\cdot}25$. Such a low threshold implies that the harm of a false-positive classification is relatively limited. In practice, it may be difficult to specify the threshold p_t exactly. A range of potential decision thresholds may hence need to be considered.

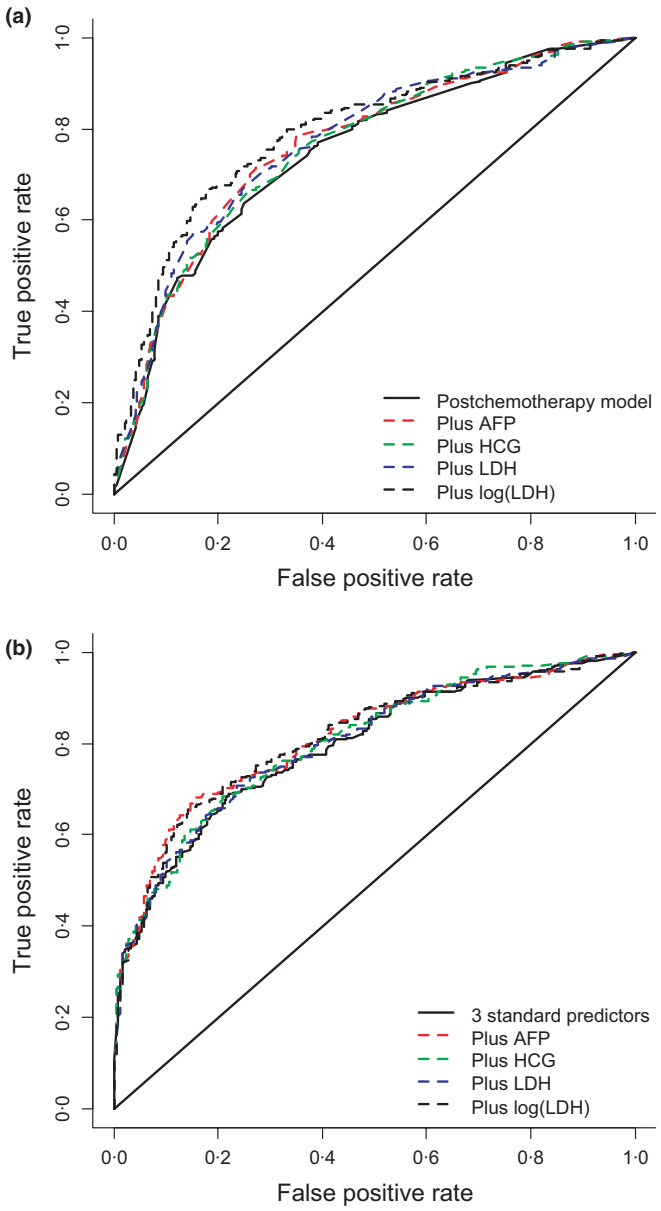

Figure 2 Receiver operating characteristic (ROC) curves for adjustment with postchemotherapy size (a) or postchemotherapy size, reduction in size, and primary histology (b).

This is carried out in a decision curve (http://www.decisioncurveanalysis.org) [27,29].

Discrepancies between NRI and ΔNB are possible when the threshold p_t is not equal to the prevalence of disease. A detailed hypothetical example is discussed in the Appendix. As a reconciliation between NRI and NB, a weighted variant of NRI has been proposed (wNRI, see Appendix). This wNRI weights the improvement in sensitivity and specificity by the consequences of TP and FP reclassifications and hence is identical to the NB except its scaling [24]. The relationship is that wNRI = $\Delta NB/p_t$. So in Table 4, wNRI is five times ΔNB.

Reclassification and NB in the case study

Reclassifications were first calculated for any change in risk estimate, i.e. a category-free version [denoted as NRI (>0), Table 4]. Compared to postchemotherapy size alone, the continuous version of LDH contributed the most, which is in agreement with the results obtained using AUC and R^2. When the contribution over a more complete reference model was studied (including size, reduction and primary histology), LDH seemed of less relevance, with lower NRI(>0) values, while AFP had the highest NRI(>0) value, again in agreement with AUC and R^2. Of note, however, NRI(>0) indicated reasonable potential for correct reclassification using HCG regardless of the baseline model, while HCG looked least important to add according to AUC or R^2.

Further analyses considered a binary classification, based on a clinically relevant threshold for the risk of tumour. This threshold was based on a previously performed formal decision analysis, where estimates from literature and from experts in the field were used to weight the harms of missing tumour against the benefits of resection in those with tumour [30]. This decision analysis indicated that a risk threshold of 20% would be clinically defendable.

With a 20% threshold, the reclassification analysis suggested that continuous LDH measurements, abnormal AFP and abnormal HCG offered reasonable improvement when added to a model with postchemotherapy size alone [NRI(0·20)s around 0·10]. The decision-analytic measure picked AFP and HCG (ΔNB +0·64% and +0·60% more TPs for the same number of FPs for abnormal AFP and HCG, respectively) as the best markers but not dichotomized or continuous LDH (ΔNB +0·14% and +0·23% more TPs for the same number of FPs for dichotomized and continuous LDH, respectively). When we considered the model with three standard predictors (postchemotherapy size, reduction and primary histology), ΔNB was only positive for adding abnormal HCG whereas NRI(0·20) suggested that both abnormal HCG and AFP might improve reclassification. The wNRI followed the exact same pattern as the NB analyses.

We note that in many instances, the differences between improvements in model performance for the three markers were relatively small (see e.g. Fig. 3) and that all differences were quite uncertain. Most 95% confidence intervals included zero for the reclassification and clinical usefulness measures at the 20% threshold [NRI(0·20), wNRI 0·20, ΔNB(0·2)], while most NRI(>0) results were statistically significant, in line with the odds ratios (Table 2) and continuous measures of improvement in model performance (ΔAUC, ΔR^2, Table 3).

Additional interesting insights can be derived by examining the components of NRI(0·20) and NRI(>0) presented in Table 4.

Table 4 Incremental value according to NRI and NB of three markers over a model including postchemotherapy size and a model including postchemotherapy size, reduction in size, and primary tumour histology

Extension	NRI(> 0)	NRI(0·2)	wNRI(0·2)	ΔNB(0·2)
Compared to postchemotherapy size				
AFP abnormal	+0·52–0·06 = +0·46 (0·30–0·61)	−0·01 + 0·11 = +0·096 (−0·01–0·20)	+0·032 (−0·02–0·08)	+0·64% (−0·34–1·6%)
HCG abnormal	+0·41–0·04 = +0·37 (0·21–0·53)	−0·01 + 0·10 = +0·092 (−0·01–0·20)	+0·030 (−0·02–0·08)	+0·60% (−0·35–1·5%)
LDH abnormal	−0·34 + 0·51 = +0·17 (−0·06–0·40)	−0·02 + 0·09 = +0·077 (−0·02–0·17)	+0·007 (−0·05–0·06)	+0·14% (−0·86–1·1%)
continuous	+0·39 + 0·11 = +0·50 (0·32–0·68)	−0·02 + 0·13 = +0·111 (0·02–0·20)	+0·011 (−0·05–0·06)	+0·23% (−0·93–1·3%)
Compared to postchemotherapy size, reduction, and primary histology model				
AFP abnormal	+0·52 –0·06 = +0·46 (0·30–0·61)	−0·03 + 0·11 = +0·080 (−0·01–0·17)	−0·021 (−0·06–0·10)	−0·41% (−1·2–2·0%)
HCG abnormal	+0·41–0·04 = +0·37 (0·21–0·53)	−0·00 + 0·08 = +0·078 (−0·01–0·17)	+0·037 (−0·03–0·10)	+0·74% (−0·6–2·0%)
LDH abnormal	−0·25 + 0·40 = +0·15 (−0·10–0·40)	−0·00–0·01 = −0·012 (−0·08–0·06)	−0·014 (−0·04–0·07)	−0·28% (−1·4–0·8%)
continuous	+0·27 + 0·04 = +0·23 (0·04–0·42)	−0·01 + 0·02 = +0·007 (−0·05–0·07)	−0·025 (−0·04–0·09)	−0·51% (−1·8–0·8%)

*NRI, net reclassification index; ΔNB, difference in Net benefit. Values between brackets are 95% confidence intervals.
Values for the NRI are the NRI for patients with tumor, for patients with benign tissue, and the sum of these numbers.
NRI(> 0) was calculated using all decision thresholds.
NRI(0·20), weighted NRI and ΔNB values are calculated at a threshold probability of 20% for presence of residual tumor.

When using NRI(0·20) with a single classification threshold at 0·20, we notice that the observed improvement in reclassification (where present) is almost exclusively because of improvements in specificity. This is in apparent contrast to the category-less NRI(>0) for which large values are driven primarily by increase in event probabilities for cases. This suggests that a different choice of threshold could offer different conclusions about the relative usefulness of the markers considered. It also helps explain the observed disagreement between ΔAUC, ΔR^2 and NRI(>0) vs. measures of improvement in clinical usefulness (wNRI and ΔNB): the continuous measures pick markers that have the greatest potential for model improvement across all potential thresholds, but this potential may not be realized for a given particular threshold that is the most clinically relevant in a particular setting (as in our example).

Discussion

Various traditional and novel approaches are available to assess the incremental value of a marker, but they led to different conclusions in a case study considering three tumour markers for patients with testicular cancer. The application to a real data set highlighted some of the challenges in the assessment of the value of a marker.

Challenges in assessing markers

An important issue is the coding of continuous marker values. In our case study of patients with testicular cancer, we found that LDH, as expected, performed better with a continuous coding than with a dichotomized coding. Next to a linear transformation, at least a logarithmic transformation should be examined in marker studies. More flexible approaches are readily available nowadays, including various variants of spline functions. Graphical illustrations will often be necessary when nonlinear relationships are modelled, making the mathematics underlying the relationships less relevant.

Note that the interpretation of an OR is straightforward for a binary marker, where the OR reflects the effect of a positive marker value vs. a negative marker value. A high OR does, however, not directly mean that a marker has high additional value, because a positive marker value may be quite rare. A marker with an OR of 2 and a 50 : 50 distribution of positive and negative values may hence be considered to be far more important for prediction than a marker with an OR of 10 and a 1 : 99 distribution of positive and negative values [31,32].

For a continuous marker, the OR may often appear to be very small when the marker has a wide range of values. Sometimes, standardized effects may be shown, i.e. the effect per standard deviation change in marker value. For example, the effect of CRP in predicting cardiovascular disease is often expressed per SD change in log(CRP) value [33]. A general approach for continuous markers is to express the effect for the interquartile range, e.g. comparing the effect for the 75 percentile vs. the 25 percentile of the marker distribution [9,11].

Next, the choice of reference model was essential when assessing predictive value. A simple model with one key diagnostic characteristic (postchemotherapy mass size) led to an overall quite positive appraisal of the value of LDH.

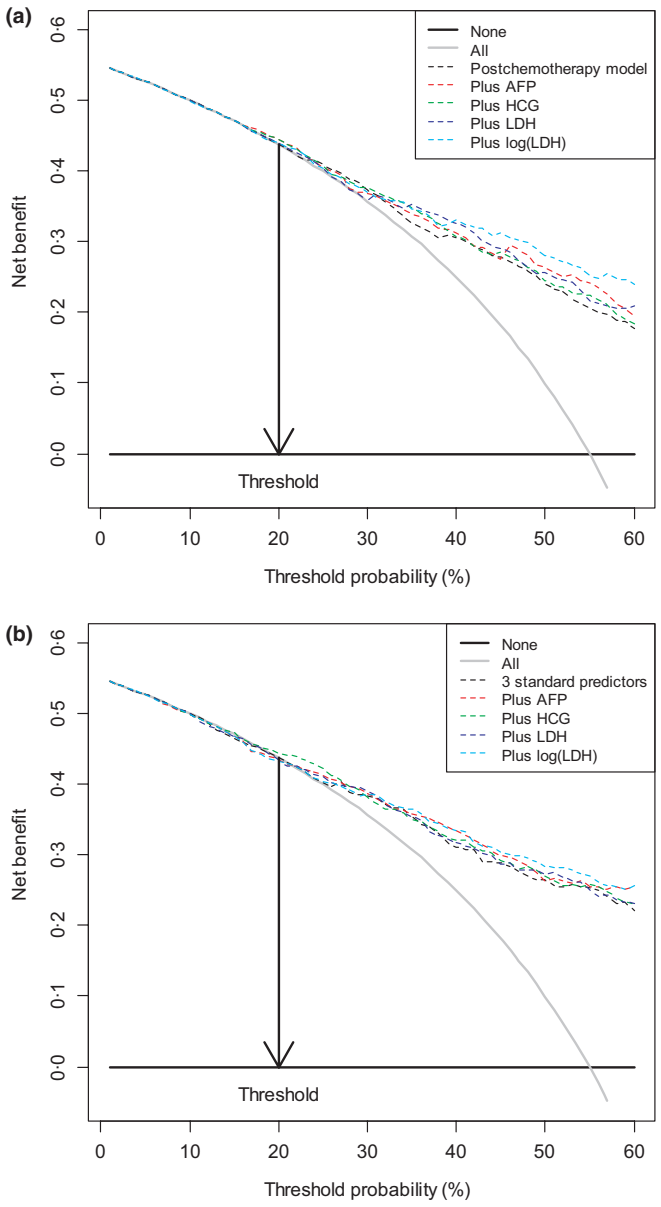

Figure 3 Decision curves showing the Net Benefit (NB) in comparison to a logistic regression model with postchemotherapy size (a) or postchemotherapy size, reduction in size, and primary histology (b).

We consider this misleading, because a full adjustment for three characteristics made that AFP or HCG looked more valuable.

Furthermore, we found consistency between the performances as judged by R^2 and the AUC (or *c* statistic) values. This may generally be expected because both consider the full distribution of predictions. Technically speaking, Nagelkerke's R^2 is a logarithmic scoring rule, and *c* a rank-order scoring rule [9]. Pearson's R^2 (or the Brier score) is a quadratic scoring rule. Each of these (Nagelkerke R^2 and *c*) led to similar conclusions on the value of a tumour marker in the case study. We could not study all overall measures of performance that have recently been proposed. These include predictiveness curves [34] and Lorenz curves [35], which are related to R^2 and AUC measures. The category-less NRI(>0) was generally consistent with R^2 and the AUC.

On the other hand, the conclusions derived using R^2 and the AUC were different from those derived using NB with the specific decision threshold of 20%. The NB analysis with the one predictor reference model suggested abnormal AFP and HCG as the best markers, whereas R^2 and the AUC indicated that continuous LDH is the most useful. When the three predictor model was used as a reference, NB favoured HCG whereas R^2 and the AUC picked AFP. Interestingly, the two reclassification measures, NRI(> 0) and NRI(0·20), fell in the middle, suggesting relatively good reclassification potential for markers picked by the R^2 and AUC as well as NB analyses. Examining event and non-event reclassification components of the NRIs offered additional valuable insights helping to explain why and how measures that integrate across all thresholds may not agree with measures that focus on one particular threshold.

We note, however, that random noise may explain a substantial part of these differences, as reflected in wide confidence intervals in Tables 3 and 4. Analyses were quite sensitive to the specific threshold chosen (results not shown), and further research should consider stable estimation of increases in NRI and ΔNB, e.g. using smoothing techniques.

Predictions vs. decisions

Which measure to use when? It is essential to realize that the main separation is between the assessment of the quality of predictions from a model vs. the assessment of the quality of decisions (or classifications) from a rule. 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 [36]. 'Prediction model' and 'prediction rule' are hence 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, R^2, category-free NRI and multiple category NRI deal with 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 of the decision [37]. A false-positive classification (overdiagnosis) is often weighted less in medical contexts than a false-negative classification (underdiagnosis of disease) [36]. In the case study, unnecessary surgery for a

Box 1 Proposal for assessing incremental value of a diagnostic test or marker

Analysis of data where the marker is studied

- Calculate difference in area under the curve (AUC) or related measures to indicate overall improvement in discrimination (AUC is a standard measure which considers the full range of potential decision thresholds).
- Calculate difference in decision-analytic performance measures, such as the net benefit or the weighted Net Reclassification Index to indicate clinical usefulness over a smaller range of medically relevant thresholds (Decision analytic measures consider the consequences of decisions explicitly).

Further studies

- Assess impact on decision making in prospective studies (If decision making is not influenced by knowledge of the marker's value, patient outcomes can not improve).
- Assess impact on patient outcome in prospective studies, preferably randomized trials, or cost-effectiveness modeling (Impact on patient outcome proves the ultimate usefulness of a marker, while finally the balance between incremental costs and incremental effects has to be considered).

benign mass should be avoided, but is less an error than withholding surgery in patient with residual tumour. The decision threshold of 20% reflects the 1:4 relative weights of these errors. 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 principle is violated in the default NRI for two categories, but followed in the wNRI [24], the NB and related measures such as the relative utility [38]. The two category NRI only is consistent with ΔNB if the decision threshold is equal to the prevalence. This is because NRI then is the sum of the improvement in sensitivity and specificity and hence implicitly weights by prevalence of disease. Further research should address the relationship between wNRI and NB in more detail.

Recommendations for marker assessment

For the evaluation of incremental value of a diagnostic or prognostic marker, the relevant comparison is between a prediction model with and without the marker. For the overall improvement in discriminative ability, the currently standard measure, the AUC or *c* statistic (Box 1), remains a valuable tool [39,40].To overcome some of its limitations [22,41], it may be useful to present increase in Nagelkerke's R^2 or the IDI as well as its 'nonparametric' version, the NRI(> 0). All these measures have their limitations if we consider a specific decision threshold, because a substantial or small increase in AUC or R^2 achieved by adding a marker to a model may not translate to substantial or small clinical usefulness at a given threshold [42]. As a next step, we therefore should consider decision-analytic measures, such as the NB, or the wNRI.

What distorts the relationship between AUC and NB? If assumptions, such as linearity of continuous predictors and additivity, hold in a logistic regression model, the ROC curve of a model with a marker is dominant to the ROC curve of a model without the marker. So, we can always find a decision threshold where both sensitivity and specificity are better in the model with the marker than the sensitivity and specificity in a model without the marker. If model assumptions are not fully fulfilled, we may have nonconcave or even crossing ROC curves. This implies that the marker is especially useful for some parts of the ROC curve. But generally speaking, a minor increase in ROC area will imply limited clinical usefulness.

Another issue is that the decision threshold may be at the outside of the distribution of predicted probabilities. This implies lower clinical usefulness compared to not using a model. This was the case for the 20% decision threshold for the risk of residual cancer. A higher threshold, closer to the prevalence of 55%, would imply much greater clinical usefulness of any of the three tumour markers considered (AFP, HCG or LDH, see Fig. 3). Generally speaking, a marker will be most clinically useful when the externally defined decision threshold is close to the prevalence of disease, that is in the middle of the risk distribution [42]. Note that the decision threshold is determined by the specific medical context and outside the influence of the modeller.

Some guidelines for marker assessment emphasize calibration [39]. Calibration refers to the agreement of predicted probabilities to observed outcome frequencies. This property of model predictions is indeed essential when we consider application of a model in a new setting to guide decision-making [9]. Calibration may, however, be less relevant when we consider the incremental value of a marker in the same data set as where we fit the reference model. Further research should address the interrelationships between measures for

Box 2 Common errors in the assessment of the value of a diagnostic test or marker

- Interpreting without considering standard predictors such as demographic and other simple characteristics (Wrong because incremental value over standard predictors is the key question).
- Dichotomizing continuous marker values (Wrong because information is lost; dichotomizaton should only be done at the end of the modeling process, for predictions that inform decision making).
- Interpreting a large odd ratio (OR) as evidence of incremental value (Wrong because OR depends on coding; and OR value ignores distribution. A high OR for a rare characteristic has limited value for diagnosing disease or predicting an outcome disease).
- Interpreting a low *P*-value as evidence of incremental value (Wrong because *P*-value depends not only on effect size but also on sample size; low *P*-values may easily be found in large studies).
- Interpreting a large value of AUC as evidence for good clinical usefulness (Wrong because AUC values can not be interpreted without context; a value of 0·7 or 0·8 may imply clinical usefulness in some settings but not in others, depending on where the decision threshold is in the distribution of predicted risks; the same holds for increases in AUC by a marker [by e.g. 0·01 or 0·02)].

discrimination, calibration and clinical usefulness. A specific issue is the challenge to find an accessible presentation and communication format for such measures to a clinical audience.

Common errors

Some errors are common in the assessment of the value of a test or marker (Box 2). Dichotomizing continuous variables is common in the epidemiological literature, while such a loss of information should be avoided [12]. As discussed, we cannot interpret a large odds ratio in a multivariable analysis as evidence of incremental value of a diagnostic marker. A high odds ratio for a rare characteristic has limited value in diagnosing disease. Another common error is to interpret a low *P*-value as evidence of incremental value. This is wrong because the *P*-value depends not only on the effect size but also on the sample size. A low *P*-value may easily be found in large studies. Instead, measures such as R^2 or the *c* statistic should be used to quantify predictive accuracy. For example, partial R^2 values were highly informative to indicate the relative importance of 26 prognostic markers of 6 month outcome in traumatic brain injury [43]. As discussed earlier, any serious evaluation of a diagnostic marker should consider a full set of standard predictors such as demographic and other simple characteristics as a reference to improve upon [3]. Also, a large increase in AUC is not sufficient evidence of good clinical usefulness, because clinical usefulness also depends on where the decision threshold is in the distribution of predicted risks [42].

Validation and impact assessment

Final points to emphasize include validation and prospective assessment of impact on clinical care (Box 1). It is common that initial studies of markers show promising results, with disappointment in later evaluations. Hence, validation in independent data is generally considered essential for confidence in the incremental value of a marker. Internal validation with cross-validation or bootstrapping is a minimum requirement [44]. Moreover, performance measures may depend on outcome definitions, the types of patients ('case-mix'), the setting and the amount of prior testing [45]. In our illustrative case study, we only showed performance in the development data, and not in independent external validation data. The relatively large sample size (n = 544, 299 with residual tumour) made that statistical optimism was small (no risk of overfitting). Moreover, external validation studies have confirmed our results [46].

Next to validation and assessment of diagnostic value, prospective impact studies need to be considered [47]. 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 treatments 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 and quality of life), with consideration of process outcomes (diagnostic tests and treatments administered) as intermediate study endpoints [39]. Because randomized trials may often not be feasible in terms of required research funding and required sample size, formal decision-analytic modelling may also be relevant [48]. In such models, we can combine estimates of the performance of the diagnostic model with and without the marker with evidence on the effectiveness of treatments that are more appropriately targeted to those who need it with a marker than without.

Conclusions

Reporting on the increase in discrimination [using ΔAUC or Δ*c* statistic, ΔR^2, IDI or NRI(> 0)] is relevant to obtain insight into the incremental value of a marker. Decision-analytic measures

such as NB or wNRI should be reported if the prediction model including the marker is to be used for making decisions. Although the standard NRI quickly gained popularity in major medical journals, researchers need to be aware of the implicit weighting of false-positive and false-negative decisions based on disease prevalence that it contains. This weighting may not be appropriate in many medical applications [49]. Hence, the components of the NRI for diseased and nondiseased subjects should always be reported, and wNRI may be considered as a better summary measure. In applications calling for a prediction rule with two categories, decision-analytic measures, such as wNRI or NB, and the corresponding decision curve, may provide the most informative metrics.

Acknowledgements

We would like to thank two anonymous reviewers for their constructive comments which helped to improve this paper. 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).

Address

Department of Public Health, Erasmus MC, Rotterdam, The Netherlands (E. W. Steyerberg, H. F. Lingsma, B. Van Calster); Department of Biostatistics, Boston University, and Harvard Clinical Research Institute, Boston, MA (M. J. Pencina); Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA (M. W. Kattan); Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA (A. J. Vickers); Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium (B. Van Calster).

Correspondence to: Ewout W. Steyerberg, PhD, Department of Public Health, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands. Tel.: 31 10 703 8470; fax: 31 10 463 8465; e-mail: e.steyerberg@erasmusmc.nl

Received 19 December 2010; accepted 26 May 2011

References

1 Ioannidis JP. Expectations, validity, and reality in omics. *J Clin Epidemiol* 2010;**63**:945–9.
2 Janssens AC, Ioannidis JP, Bedrosian S, Boffetta P, Dolan SM, Dowling N *et al*. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. *Eur J Clin Invest* 2011;doi: 10.1111/j.1365-2362.2011.02493.x. [Epub ahead of print].
3 Kattan MW. Judging new markers by their ability to improve predictive accuracy. *J Natl Cancer Inst* 2003;**95**:634–5.
4 McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 2005;**97**:1180–4.
5 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.
6 Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;**302**:2345–52.
7 Steyerberg EW, Keizer HJ, Stoter G, Habbema JD. Predictors of residual mass histology following chemotherapy for metastatic non-seminomatous testicular cancer: a quantitative overview of 996 resections. *Eur J Cancer* 1994;**30A**:1231–9.
8 Steyerberg EW, Keizer HJ, Fossa SD, Sleijfer DT, Toner GC, Schraffordt Koops H *et al*. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. *J Clin Oncol* 1995;**13**:1177–87.
9 Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer, 2009; **XXVIII**:500 p.
10 Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. *BMJ* 2002;**324**:477–80.
11 Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer, 2001;**xxii**:568 p.
12 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;**25**: 127–41.
13 Royston P, Sauerbrei W. *Multivariable Model-Building : A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables*. Chichester, England; Hoboken, NJ: John Wiley, 2008;**xvii**:303 p.
14 Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001;**xvi**:533 p.
15 Harrell FE Jr, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *J Natl Cancer Inst* 1988;**80**:1198–202.
16 Gail MH. The estimation and use of absolute risk for weighing the risks and benefits of selective estrogen receptor modulators for preventing breast cancer. *Ann N Y Acad Sci* 2001;**949**:286–91.
17 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.
18 Vittinghoff E. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer, 2005;**xv**:340 p.
19 Mittlbock M, Schemper M. Explained variation for logistic regression. *Stat Med* 1996;**15**:1987–97.
20 Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika* 1991;**78**:691–2.
21 Altman DG. ROC curves and confidence intervals: getting them right. *Heart* 2000;**83**:236.
22 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;**115**:928–35.
23 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. discussion 207–112.
24 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.

25 Pepe MS, Feng Z, Gu JW. 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**:173–81.
26 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Comments on 'Integrated discrimination and net reclassification improvements-Practical advice'. *Stat Med* 2008;**27**:207–12.
27 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565–74.
28 Peirce CS. The numerical measure of success of predictions. *Science* 1884;**4**:453–4.
29 Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. *Med Decis Making* 2008;**28**:146–9.
30 Steyerberg EW, Marshall PB, Keizer HJ, Habbema JD. Resection of small, residual retroperitoneal masses after chemotherapy for nonseminomatous testicular cancer: a decision analysis. *Cancer* 1999;**85**:1331–41.
31 McHugh GS, Butcher I, Steyerberg EW, Lu J, Mushkudiani N, Marmarou A *et al*. Statistical approaches to the univariate prognostic analysis of the IMPACT database on traumatic brain injury. *J Neurotrauma* 2007;**24**:251–8.
32 Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* 2006;**8**:395–400.
33 Shah T, Casas JP, Cooper JA, Tzoulaki I, Sofat R, McCormack V *et al*. Critical appraisal of CRP measurement for the prediction of coronary heart disease events: new data and systematic review of 31 prospective cohorts. *Int J Epidemiol* 2009;**38**:217–31.
34 Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM *et al*. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol* 2008;**167**:362–8.
35 Gail MH. Applying the Lorenz curve to disease risk to optimize health benefits under cost constraints. *Stat Interface* 2009;**2**:117–21.
36 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.
37 Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;**302**:1109–17.
38 Baker SG. Putting risk prediction in perspective: relative utility curves. *J Natl Cancer Inst* 2009;**101**:1538–42.
39 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.
40 Levinson SS. Clinical validation of biomarkers for predicting risk. *Adv Clin Chem* 2009;**48**:1–25.
41 Pepe MS, Janes H, Gu JW. Letter by Pepe *et al*. regarding article, ''Use and misuse of the receiver operating characteristic curve in risk prediction''. *Circulation* 2007;**116**:e132. author reply e134.
42 Coppus SF, van der Veen F, Opmeer BC, Mol BW, Bossuyt PM. Evaluating prediction models in reproductive medicine. *Hum Reprod* 2009;**24**:1774–8.
43 Murray GD, Butcher I, McHugh GS, Lu J, Mushkudiani NA, Maas AI *et al*. Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007;**24**:329–37.
44 Steyerberg EW, Harrell FE Jr, 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.
45 Bossuyt PM. Clinical validity: defining biomarker performance. *Scand J Clin Lab Invest Suppl* 2010;**242**:46–52.
46 Vergouwe Y, Steyerberg EW, Foster RS, Sleijfer DT, Fossa SD, Gerl A *et al*. Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: a model update and multicentre validation with more than 1000 patients. *Eur Urol* 2007;**51**:424–32.
47 Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;**338**:b606.
48 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.
49 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' by M. J. Pencina *et al*., Statistics in Medicine. *Stat Med* 2008; **27**:199–206. doi: 10.1002/sim.2929.

Appendix

Relationship between NB and NRI, leading to a weighted NRI (wNRI).

Hypothetical example to illustrate discrepancy between NRI and NB

Consider 1000 patients, 500 with and 500 without disease. Marker A correctly reclassifies 100 subjects without disease and falsely reclassifies 50 subjects with disease. Marker B falsely reclassifies 100 subjects without disease and correctly reclassifies 50 subjects with disease. The NRI for marker A is the sum of the improvements in sensitivity and specificity: –50/500 + 100/500 = +0·10. In contrast, the NRI for marker B is +50/500 −100/500 = −0·10. If the decision threshold is 20%, we should, however, weight the FP reclassifications as 0·25 times a TP reclassification. Hence, the differences in NB are (−50 + 0·25*100)/1000 = −0·025 for marker A and (50 − 0·25*100)/1000 = +0·025 for marker B. Hence, NRI and ΔNB have opposite directions in this example. The NB calculation recognizes that marker B is more clinically useful because 50 more TP reclassifications outweigh the 100 more FP reclassifications.

Notation for further derivation of interrelationship

We assume a data set of size N, with N_+ diseased and N_- nondiseased subjects such that $N_+ + N_- = N$. The prevalence of the disease is denoted as P, and the probability threshold to triage patients as low or high risk as p_t. Using p_t, TP represents the number of TPs (diseased patients predicted to be at high risk), FP the number of FPs (nondiseased patients predicted to be at high risk), TN the number of true negatives (nondiseased patients predicted to be at low risk) and FN the number of false negatives (diseased patients predicted to be at low risk).

If we have two diagnostic prediction models, one with standard predictors (model 1) and one with standard predictors and new diagnostic marker (model 2), the TPs for these models, for example, are denoted by TP_1 and TP_2, respectively.

Net reclassification improvement

The NRI is computed as the sum of differences in proportions of individuals moving up minus the proportion moving down for those with the outcome, and the proportion of individuals moving down minus the proportion moving up for those without the outcome. In case of a single cut-off, moving up means that adding the marker changes the prediction from low to high risk while moving down implies an opposite reclassification. Following Pencina *et al.* [23], the NRI is given as

$$\begin{aligned}\mathrm{NRI} = {} & P(\mathrm{up}|\mathrm{diseased}) - P(\mathrm{down}|\mathrm{diseased}) \\ & + P(\mathrm{down}|\mathrm{non-diseased}) - P(\mathrm{up}|\mathrm{non-diseased}).\end{aligned}$$

For binary classification as low or high risk, this reduces to the sum of the improvements in sensitivity and specificity, and the formula can be written as

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

Thus, the NRI implicitly weights TP and FP improvements by prevalence even though p_t conveys information about misclassification costs.

Net benefit

The NB is a measure that explicitly incorporates weights for detecting disease (TP) vs. overdiagnosing nondisease (FP). The NB can be interpreted as the fraction of TP classifications penalized for FP classifications, and its formula is

$$\mathrm{NB} = \frac{\mathrm{TP}}{N} - w\frac{\mathrm{FP}}{N}, \quad \text{with } w = \frac{p_t}{1-p_t}.$$

This shows that NRI is consistent with the decision-analytic NB only if $p_t = P$. Else, NRI uses weights that differ from the misclassification costs implicitly assumed through p_t.

weighted NRI

Using Bayes' rule, the original formulation of the NRI can be rewritten [24]:

$$\begin{aligned}\mathrm{NRI} = {} & \frac{P(\mathrm{diseased}|\mathrm{up})P(\mathrm{up}) - P(\mathrm{diseased}|\mathrm{down})P(\mathrm{down})}{P(\mathrm{diseased})} \\ & + \frac{P(\mathrm{non\text{-}diseased}|\mathrm{down})P(\mathrm{down}) - P(\mathrm{non\text{-}diseased}|\mathrm{up})P(\mathrm{up})}{P(\mathrm{non\text{-}diseased})}.\end{aligned}$$

We denote the benefit when a diseased patient is reclassified upwards by model 2 relative to model 1 by s_1. Likewise, s_2 is used to denote the benefit obtained when a nondiseased patient is reclassified downwards. The weighted NRI, wNRI [24], equals

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

For binary classification, this can be reduced to

$$\mathrm{wNRI} = s_1\frac{\mathrm{TP}_2 - \mathrm{TP}_1}{N} + s_2\frac{\mathrm{FP}_1 - \mathrm{FP}_2}{N}.$$

The default values for the weights s_1 and s_2 are $\frac{1}{P}$ and $\frac{1}{1-P}$, respectively, which reduces wNRI to the NRI. However, a decision-analytic perspective calls for weights based on p_t [37]. For example, if p_t is 0·20, it is implied that detecting disease is considered four times more important than detecting nondisease. The definition of NRI implies that the harmonic mean of s_1 and s_2 is 2. Hence, s_1 might be set to 5 and s_2 to 1·25 in this example [24].