Graphical assessment of incremental value of novel markers in prediction models: From statistical to decision analytical perspectives

Ewout W. Steyerberg[*,1], **Moniek M. Vedder**[1], **Maarten J. G. Leening**[2,3], **Douwe Postmus**[4], **Ralph B. D'Agostino Sr.**[5], **Ben Van Calster**[1,6], and **Michael J. Pencina**[7]

[1] Department of Public Health, Erasmus MC: University Medical Center Rotterdam, Rotterdam, The Netherlands
[2] Department of Epidemiology, Erasmus MC: University Medical Center Rotterdam, Rotterdam, The Netherlands
[3] Department of Cardiology, Erasmus MC: University Medical Center Rotterdam, Rotterdam, The Netherlands
[4] Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[5] Framingham Heart Study, Framingham, MA, USA
[6] Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[7] Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, Durham, NC, USA

Received 5 November 2013; revised 24 April 2014; accepted 25 May 2014

New markers may improve prediction of diagnostic and prognostic outcomes. We aimed to review options for graphical display and summary measures to assess the predictive value of markers over standard, readily available predictors. We illustrated various approaches using previously published data on 3264 participants from the Framingham Heart Study, where 183 developed coronary heart disease (10-year risk 5.6%). We considered performance measures for the incremental value of adding HDL cholesterol to a prediction model. An initial assessment may consider statistical significance (HR = 0.65, 95% confidence interval 0.53 to 0.80; likelihood ratio $p < 0.001$), and distributions of predicted risks (densities or box plots) with various summary measures. A range of decision thresholds is considered in predictiveness and receiver operating characteristic curves, where the area under the curve (AUC) increased from 0.762 to 0.774 by adding HDL. We can furthermore focus on reclassification of participants with and without an event in a reclassification graph, with the continuous net reclassification improvement (NRI) as a summary measure. When we focus on one particular decision threshold, the changes in sensitivity and specificity are central. We propose a net reclassification risk graph, which allows us to focus on the number of reclassified persons and their event rates. Summary measures include the binary AUC, the two-category NRI, and decision analytic variants such as the net benefit (NB). Various graphs and summary measures can be used to assess the incremental predictive value of a marker. Important insights for impact on decision making are provided by a simple graph for the net reclassification risk.

Keywords: Decision analysis; Reclassification; Regression analysis; Risk assessment; ROC curve.

Additional supporting information may be found in the online version of this article at the publisher's web-site

[*] Corresponding author: e-mail: e.steyerberg@erasmusmc.nl

1 Introduction

Risk prediction models, or clinical prediction models, have been successfully developed to aid clinicians in personalized decision making in all major fields of modern medicine, including cardiovascular disease, diabetes, trauma, and cancer (Steyerberg et al., 2013). Prediction models most often consider binary events, which are already present in a patient (disease, i.e. diagnosis), or occur in the future (events, i.e. prognosis). Probabilistic models for such outcomes may be evaluated with various performance measures, commonly related to discrimination (i.e. the ability to distinguish subjects with the event of interest from those without) and calibration (i.e. the agreement between predicted and observed probabilities of the event) (Harrell, 2001; Steyerberg, 2009).

Nowadays, specific interest focuses on ways in which risk prediction can be improved using novel markers (Pencina et al., 2008; Pencina et al., 2010; Pencina et al., 2011; Pencina et al., 2012b). Markers may include information from simple demographics or blood markers to genomics, proteomics, and advanced imaging techniques. Such markers hold the promise of bringing personalized medicine closer. An important question is how to evaluate the usefulness of a new marker in making better decisions in clinical practice, such as better targeting of statin therapy to those at increased risk of cardiovascular disease (Ridker et al., 2007). Various measures have been proposed to quantify usefulness, including the net reclassification improvement (NRI) (Pencina et al., 2008), and decision analytic variants such as the net benefit (NB) (Vickers and Elkin, 2006) and relative utility (RU) (Baker et al., 2009). The latter measures have important theoretical advantages (Van Calster et al., 2013). They directly translate to patient-centered outcomes, but are more difficult to communicate to a clinical audience (Localio and Goodman, 2012; Leening and Steyerberg, 2013). Graphical displays may be of assistance in assessing usefulness of a marker for predictive purposes, since graphs allow for direct pattern recognition in addition to a table look-up function (Cleveland, 1985). We illustrate various graphical possibilities and summary measures using previously published data from the Framingham Heart Study, where we focus on adding high-density lipoprotein (HDL) cholesterol as a marker to improve predictions of 10-year risk of coronary heart disease (CHD). After describing the data and the statistical analysis, we consider performance measures related to continuous predictions and to dichotomizations to provide a classification as low versus high risk.

2 Data

We provide an illustration with data published previously (Pencina et al., 2008), relating to 3264 participants from the Framingham Heart Study aged 30–74 years. Participants with prevalent cardiovascular disease and missing standard risk factors were excluded (Pencina et al., 2008). Participants were followed for 10 years for the development of a first CHD event (including myocardial infarction, angina pectoris, coronary insufficiency, or CHD death). A total of 183 individuals developed CHD (5.6% 10-year cumulative incidence).

3 Analysis

We focus on the improvement in model performance due to the addition of HDL cholesterol to a model that already contains sex, diabetes, and smoking as dichotomous predictors and age, systolic blood pressure, and total cholesterol as continuous predictors. Adding HDL cholesterol as a continuous predictor to a Cox regression model was highly significant ($\mathrm{HR} = 0.65$ per SD increase, 95% confidence interval 0.53 to 0.80, p-value < 0.001). We compare two sets of predicted probabilities of the 10-year CHD risk: one set of predictions based on a model *without* and one set of predictions based on a model *with* HDL cholesterol included. We consider the full risk distributions as well as categorization by the clinically motivated threshold of 20% 10-year CHD risk. Since we focus on conceptual issues for binary classification, we simply stratify subjects as those with events versus those without events after 10-year follow-up. We recognize that survival data may require more sophisticated approaches

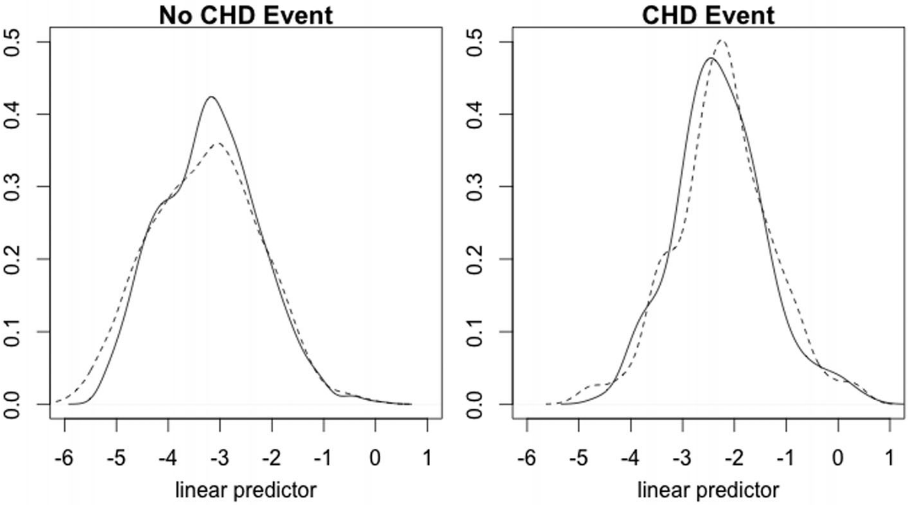

Figure 1 Density plots of log odds of predicted 10-year CHD risks (“linear predictor”) for nonevents and events in 3264 participants from the Framingham Heart Study. Solid line: distribution of linear predictor without HDL cholesterol; dotted line: distribution of linear predictor with HDL cholesterol.

to address censoring of incomplete observations, such as the use of Kaplan-Meier or Cox regression analysis to estimate expected events rather than naively use observed events (Steyerberg and Pencina, 2010; Pencina et al., 2011).

4 Performance measures related to continuous predictions

4.1 Densities and box plots with their summary measures

A better prediction model will provide predicted risks closer to the observed outcome: higher predicted risks for those with an event, and lower predicted risks for those without an event. Potential graphical illustrations include a density plot (Fig. 1) or a box plot (Fig. 2), where we can compare the predicted risks from models with and without the marker (Royston and Altman, 2010). In both plots, we may choose to show predicted values transformed to the linear predictor scale (log hazard for survival, log odds for binary predictions), or at the absolute risk scale (as a percentage). In both plots, we hope to see more separation in predicted risks, i.e. lower risk predictions for those without and higher risk predictions for those with events.

A density plot may most naturally use the linear predictor scale, since a reasonably normal distribution is commonly noted at this scale (Fig. 1). A density plot for the linear predictors relates to summary measures that consider the log-based distances of predictions to observed outcomes. One such summary measure is the explained variation as defined by Nagelkerke, where the log-likelihood of a model is scaled between 0 and 100%:

$$R^2 = (1 - \exp(-\mathrm{LR}/n))/(1 - \exp(-2\mathrm{LL0}/n)),$$

where LR is the likelihood ratio statistic of the model, $-2\mathrm{LL}_0$ is the -2 log-likelihood of the Null model without any covariates, and n is the number of subjects (Nagelkerke, 1991). In our example, Nagelkerke’s R^2 increased from 13.1% to 14.7% or +1.6%.

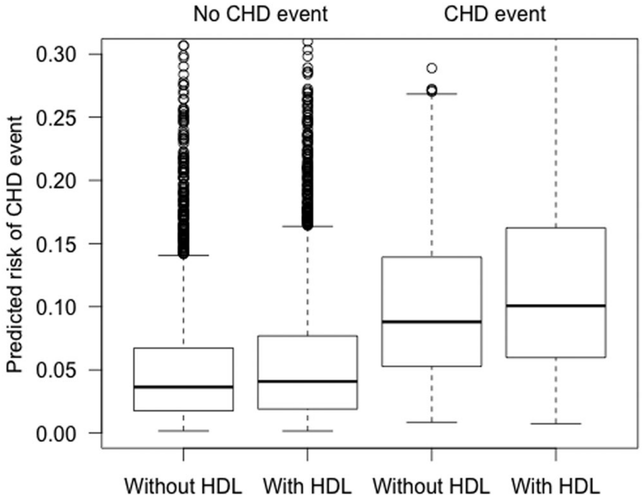

Figure 2 Box plots of predicted 10-year CHD risks for patients without and with a CHD event in 3264 participants from the Framingham Heart Study for the model without HDL and with HDL.

For box plots, the use of absolute risks has been proposed before. Box plots are attractive as a visual companion to measures such as the discrimination slope, which is defined as the difference in mean predicted risks for those with and without the event. The discrimination slope increased from 6.29% to 7.14% (+0.85%), which is identical to another relatively novel measure: the integrated discrimination index (IDI, +0.0085) (Pencina et al., 2008). The difference in Pearson's R^2 (defined simply by comparing predicted risks to observed outcomes) was +1.9% and hence larger than the IDI, although these estimates should be asymptotically equivalent (Tjur, 2009).

4.2 Predictiveness curves

The absolute risk can also be shown as a cumulative distribution in a predictiveness curve, i.e. based on the ordering of risk from lowest to highest values (Pepe et al., 2008). Better predictiveness is indicated by more spread in predictions, with many low predictions and a steep increase in the cumulative distribution of predictions at 1 minus the event rate. The event rate is useful as a reference line, with nonpredictive model predictions being close to that line. In Fig. 3, we note that the distribution of the risks based on the model with HDL is only slightly more extreme than the distribution based on the model without HDL, consistent with the minor increase in R^2 statistics as noted above. The performance of the models in the population is illustrated with a 20% risk threshold, where we note that both models would classify approximately 96% as low risk and 4% as high risk.

As for Fig. 1 and Fig. 2, we can stratify the predictiveness curve by event status. The specificity is related to the cumulative distribution of risks for those with no CHD event. Specificities were very similar in Fig. 3, but the sensitivity of predictions from a model with HDL was higher than from a model without, consistent with Fig. 1 and Fig. 2.

4.3 ROC curves and area under the curve

The receiver operating characteristic (ROC) curve shows the relation between sensitivity (or the true positive rate, i.e. the fraction with a predicted probability above a cut-off among those developing CHD), versus 1 minus specificity (or the false positive rate, i.e. the fraction with a predicted probability

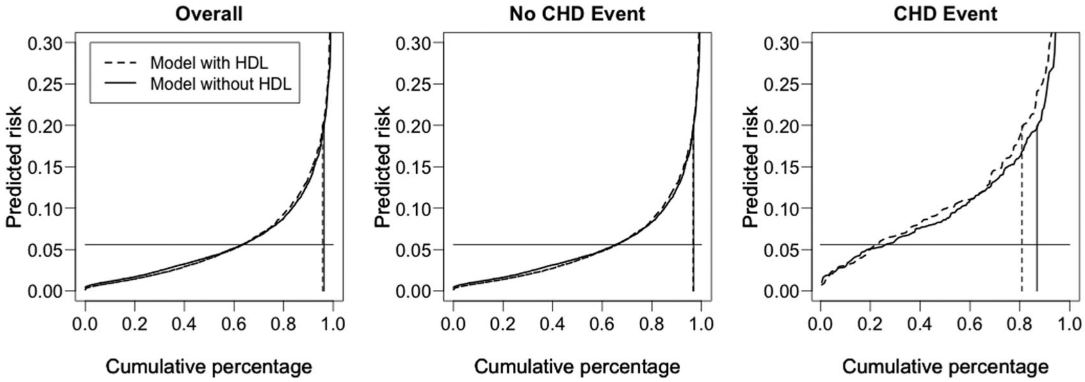

Figure 3 Predictiveness curves for CHD events in 3264 participants from the Framingham Heart Study. The horizontal lines indicate the event rate (5.6%). The vertical lines indicate the 20% decision threshold, which leads to similar specificity (96.82% and 96.66%) but higher sensitivity (13.1% and 19.1%) for the models without and with HDL cholesterol, respectively.

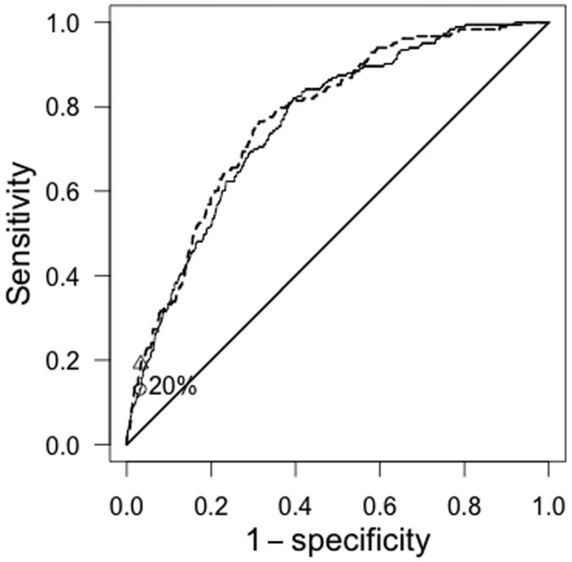

Figure 4 Receiver operating characteristic (ROC) curves. Solid lines: based on 10-year CHD predictions without HDL cholesterol; dotted lines: based on 10-year CHD predictions with HDL cholesterol in 3264 participants from the Framingham Heart Study. The 20% decision threshold is indicated with a triangle and a circle for the model with and without HDL cholesterol, respectively.

above a cut-off among those not developing CHD, Fig. 4). The sensitivity and specificity pairs are calculated for consecutive cut-offs for the predicted probabilities of the 10-year CHD risk. The area under the ROC curve (AUC) is the most popular metric to quantify discriminative ability, i.e. the ability to distinguish those who will develop the event of interest from those who will not. The AUC is the probability that given two subjects (one with CHD and one without CHD), the model will assign a higher probability of CHD to the former. The AUC is a rank order statistic, and when estimated nonparametrically, related to the Mann–Whitney U statistic: AUC = U/(n1 × n2), where n1 and n2

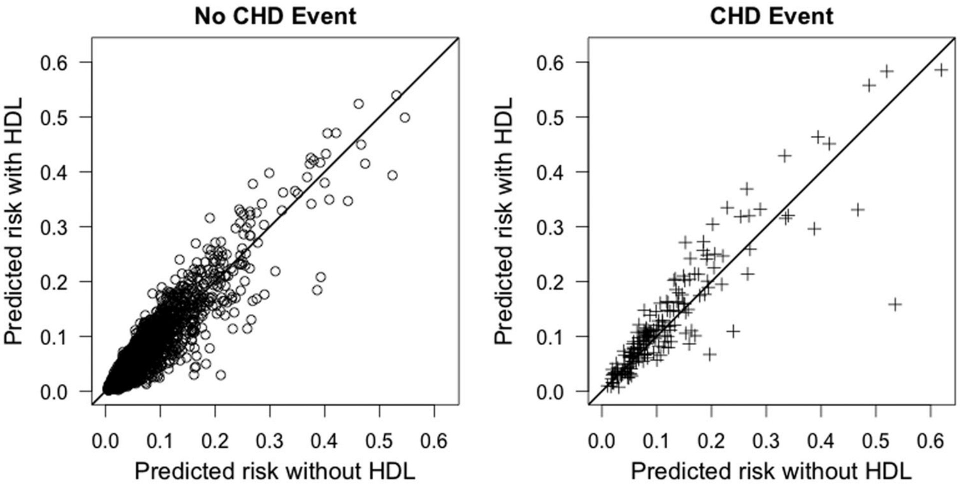

Figure 5 Reclassification graphs for the addition of HDL cholesterol to 10-year CHD predictions in 3264 participants from the Framingham Heart Study. Classification with a 20% risk cut-off is indicated with dotted lines.

are the numbers of persons with and without events (for which the product equals the number of possible pairs) (Hanley and McNeil, 1982). The AUC [95% confidence interval] for the model without versus with HDL was 0.762 [0.730–0.794] versus 0.774 [0.742–0.806]. We note that plotting ROC curves makes sense only when clinically relevant thresholds are indicated, at which sensitivity and specificity can be determined from the graph. The link between threshold and sensitivity/specificity is immediately clear in the predictiveness curve with stratification by event status (Fig. 3, middle and right panels). Furthermore, it should be noted that delta AUC depends strongly on the value of the reference model without the marker (Pencina et al., 2012a; Austin and Steyerberg, 2013).

4.4 Reclassification plots and the Net Reclassification Improvement

A simple and informative graphical summary can be obtained by plotting the predicted probabilities based on the model with versus that without the new marker(s) with different symbols denoting subjects with and without events (Fig. 5) (McGeechan et al., 2008). If the new marker offers no improvement, the points will scatter around the diagonal line. If the new model is useful, the predicted probabilities for events will be larger using the new model and hence points denoting events will lie above the diagonal line. Similarly, predicted probabilities for nonevents will be smaller using the new model and hence points denoting nonevents will lie below the diagonal line. The extent to which a clear separation can be seen helps determine the degree of model improvement. A limitation is that for large data sets, the graphical impression may be difficult to notice with many overlapping points (McGeechan et al., 2008).

This graphical presentation can be summarized using the continuous net reclassification improvement (NRI($>$0)) (Pencina et al., 2011). In Fig. 5, 62.3% among those with events had a predicted risk with HDL that was higher than the predicted risk without HDL, i.e. a better risk estimate. Conversely, 37.7% had a predicted risk with HDL that was lower than the predicted risk without HDL. The net proportion with better risk estimates (or event NRI($>$0)) hence was 24.6%. Similarly, we note that 52.8% of those without an event had a lower prediction with HDL than without, while 47.2% had a higher prediction (nonevent NRI($>$0) $=$ 5.5%). The total continuous NRI($>$0) equals the sum of the two components above, $0.246 + 0.055 = 0.301$.

It has been noted that NRI(>0) can be viewed as a measure of effect size of the new predictor in the context of a risk prediction model rather than a difference in performance of two models (Pencina et al., 2012a). A feature of NRI(>0) is that it is only weakly related to the performance of the reference model. Under normality, in case of an independent predictor added to the risk prediction model, it is the same no matter how good or bad the reference model is (Pencina et al., 2012a). Finally, it is important to note that the NRI(>0) considers any change in predicted risk, irrespective of the magnitude. The size of the risk difference is considered in measures such as the IDI (Pencina et al., 2008). An overview of the discussed performance measures and options for display is shown in Table 1.

5 Performance measures related to binary classification

5.1 Sensitivity and specificity

When we focus on one particular decision threshold, such as 20% risk of CHD, the changes in sensitivity and specificity are central to quantify the incremental value of a marker (Van Calster et al., 2013). As a first option for graphical display, we indicated the 20% risk threshold in the predictiveness and ROC curves (Figs. 3 and 4). At the 20% threshold, the sensitivity increased from 13.1% to 19.1%, and specificity decreased from 96.8% to 96.7% with the addition of HDL cholesterol to the model. The increase in sensitivity and decrease in specificity is in line with theoretical expectations, since the decision threshold was higher than the event rate (Van Calster et al., 2014).

5.2 Net Reclassification Risk graph

For a direct understanding of the clinical usefulness of a marker at a specific decision threshold, we propose a "net reclassification risk" graph. This graph allows us to focus on the number of reclassified individuals and their observed event rates. The components for this graph are identical to what is used in a reclassification table. In a reclassification table (Table 2), we stratify by event status to calculate net improvements for those with and without events. Alternatively, we may stratify by reclassification group to calculate numbers of persons and absolute risks (a net reclassification risk table, Table 3). Here, the two most relevant groups are those reclassified from low to high and from high to low risk (low: <20% risk, high: > = 20% risk). The graphical display as in Fig. 6 allows for a straightforward interpretation: a larger fraction reclassified is better (width of the bars) and more separation of absolute risks is better (vertical spread). The area of each bar is proportional to the number of events in a reclassification group. The change in sensitivity is proportional to the difference between areas for the L→H and H→L groups in Fig. 6. A marker that can better identify events hence has a larger difference in areas. From Table 3 and Fig. 6 we learn that 29 persons were reclassified from high to low risk (H->L), with a 10% event rate, and 45 reclassified from low to high risk (L→H), with a 31% event rate. We hence identify $45 \times 0.31 - 29 \times 0.10 = 11$ more events. These 11 events account for a sensitivity increase of 11/183 (6.0%). The increase is partly achieved by defining a larger number of persons as high risk (45 versus 29 persons), which causes some overtreatment for those without events (more false-positives). The number of false-positives can be calculated from the probabilities of a nonevent, which is the complement of the event risk: $45\ (1 - 0.31) - 29\ (1 - 0.10) = 5$ additional false-positives. This loss in specificity (5/3081, −0.2%) might be shown similarly in a net reclassification risk graph, either by stacked bars, or by plotting the nonevent rate at the y-axis (results not shown).

5.3 Summary measures for binary classification

The AUC for a binary ROC curve equals (sensitivity + specificity)/2. It increased from 0.550 to 0.579 (+0.029). In the case of a single cut-off, the event NRI and nonevent NRI are the changes in sensitivity and specificity. Hence,

Table 1 Overview of performance measures with display options and summary measures of performance for predicting 10-year CHD risks in 3264 participants from the Framingham Heart Study. We first consider continuous predictions, followed by dichotomized classifications. Performance relates to the difference between a prediction model with and without HDL.

Performance measure	Display option	Performance	Comments
Continuous predictions			
Δ log-likelihood	Fig. 1: Density plot	+19	We note a statistically significant improvement in model fit; 1.6% more variability is explained at the log-likelihood scale
ΔNagelkerke R^2		+1.6%	
Δlp events		+0.058	
Δlp nonevents		−0.084	
Δpredicted risk events	Fig. 2: Box plot	+0.81%	Those with events receive higher predicted risks and may hence better be identified; 1.9% more variability is explained at a squared distance scale for y vs $\hat{\mathbf{y}}$.
Δpredicted risk nonevents		−0.04%	
Δdiscrimination slope (= IDI)	Fig. 3: Predictiveness curves	+0.85%	
ΔPearson R^2		+1.9%	
ΔAUC	Fig. 4: ROC curve	0.012	The probability of correctly identifying who will develop a CHD event increases by 1.2% among a random pair of participants where 1 has an event and 1 has no event.
continuous NRI events	Fig. 5: Reclassification graph	24.6%	A net 24.6% of those with events receive higher predicted risks, and a net 5.5% of nonevents receive lower predicted risks. Their sum is 0.301.
continuous NRI nonevents		5.5%	
continuous NRI		0.301	
Dichotomized classification			
ΔAUC	Fig. 4: ROC curve	0.029	The probability of correctly identifying who will develop a CHD event increases by 2.9% among a random pair of participants where 1 has an event and 1 has no event, if a decision threshold of 20% 10-year CHD risk is applied.
NRI events	Table 2 and Table 3: Reclassification tables Fig. 3: Predictiveness curve Fig. 6: Net Reclassification Risk graph	6.0%	A net 6.0% increase in high risk classifications for those with events, with a minor net decrease (−0.2%) in low risk classifications for those without events. The sum is the NRI (0.058), which is twice the increase in AUC for this binary classification.
NRI nonevents		−0.2%	
NRI		0.058	

Table 1 Continued.

Performance measure	Display option	Performance	Comments
Net Benefit Test threshold	Table 2 and Table 3: Reclassification table Fig. 7: Decision curve as a sensitivity analysis.	3/1000 335	The net number of true positive classifications increases by 3 in 1000 where HDL cholesterol is measured, on a scale from 0 to 5.6% (event rate). This implies that 335 measurements of HDL cholesterol have to be done to identify one more event as high risk without decreasing low risk classifications for those without events.

Table 2 Reclassification table with 2 categories based on a >20% 10-year CHD risk threshold to define low and high risk.

		Model with HDL cholesterol	
		Low	High
Without event: n = 3081			
Model without HDL cholesterol	Low	2952	31
	High	26	72
With event: n = 183			
Model without HDL cholesterol	Low	145	14
	High	3	21

Table 3 Risk table for reclassified persons with the addition of HDL cholesterol, based on a >20% 10-year CHD risk threshold to define low and high risk.

Reclassification	Events	Nonevents	N	Risk [95% CI]
Low->Low	145	2952	3097	4.7% [4.0–5.5%]
High->Low	3	26	29	10.3% [2.7–28%]
Low->High	14	31	45	31.1% [19–47%]
High->High	21	72	93	22.6% [15–33%]
Total	183	3081	3264	5.6%

95% CI: 95% confidence interval

NRI, ΔNB, and ΔRU calculations from Table 2 and Table 3:

NRI events = (14 – 3)/183 = 6.0%

NRI nonevents = (26 – 31)/3081 = −0.2%

NRI = [(14 – 3)/183] + [(26–31)/3081] = 6.0% + −0.2% = 0.058

ΔNB = (11 – 0.25 × 5)/3264 = 0.003

ΔRU = (11 – 0.25 × 5)/183 = 0.053

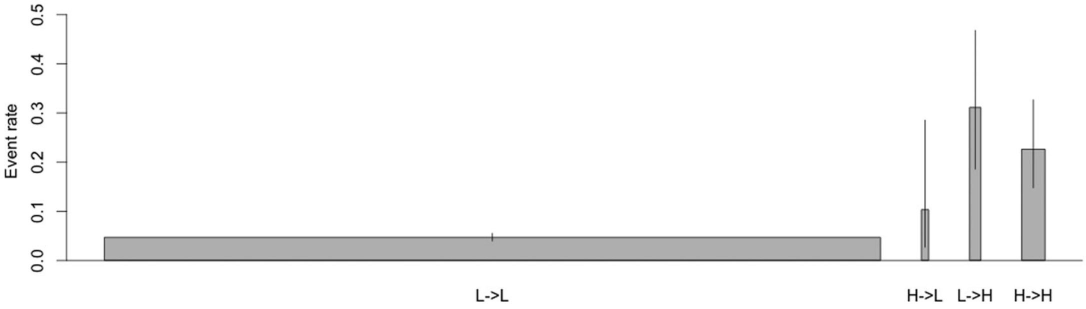

Figure 6 Net reclassification risk graph for the addition of HDL cholesterol to 10-year CHD predictions in 3264 participants from the Framingham Heart Study. L: Low risk classification; H: high risk classification, based on a 20% risk cut-off. Reclassified patients are in the groups H→L and L→H (low risk reclassified as high risk, and high risk reclassified as low risk, respectively). Uncertainty is indicated by 95% confidence intervals.

NRI = 2 × **Δ**AUC. (Pencina et al., 2008)
In our case study, NRI = 0.058 [= 6.0% − 0.2%], and **Δ**AUC = 0.029 [= 0.579 − 0.550].

ΔAUC and NRI weight changes in sensitivity and specificity equally. If the event rate is below 50%, this implies that a change in classification for an individual with an event is weighted relatively heavier than for an individual without an event. For event rates over 50%, specificity changes are weighted as relatively more important. Specifically, **Δ**AUC and NRI weight the numbers of extra TP and TN classifications by the odds of the event rate (Van Calster et al., 2013).

5.4 Weighted sums of sensitivity and specificity

A weighted variant of summing sensitivity and specificity was already proposed by Peirce in 1884 (Peirce, 1884). It was recently reintroduced as the Net Benefit (NB) (Vickers and Elkin, 2006). For changes in the number of true positives (TP) and false positives (FP), improvement in NB is defined as

$$\Delta\mathrm{NB} = (\Delta\mathrm{TP} - w \times \Delta\mathrm{FP})/\text{number of subjects}$$

where the weight w is the odds of the decision threshold. For example, a 20% CHD risk threshold means $w = 0.20/(1 - 0.20) = 0.25$. The decision threshold of 20% hence implies that we weight a FP classification 0.25 times as important as a TP classification, or that 1 more TP classification is worth four more FP classifications. In Table 3 and Fig. 6 we note 11 more TP at the price of five more FP classifications. The burden of overtreatment of those without events is explicitly weighted in the NB calculation by the odds of the decision threshold 0.20/(1 – 0.20) = 0.25. This leads to a penalty for overtreatment of 5 × 0.25 = 1.25. The **Δ**NB hence is (11 – 1.25)/3264 = 0.3%, equivalent to potentially identifying and treating an additional three events per 1000 persons screened without extra overtreatment after the addition of HDL cholesterol to the CHD prediction model.

The link between a decision threshold and the relative weight of TP versus FP classifications has a strong foundation in decision theory (Pauker and Kassirer, 1980). It is also used in other recently proposed weighted summary measures such as the change in relative utility (**Δ**RU) (Baker, 2009) and the weighted NRI (wNRI) (Pencina et al., 2011). In our example, the decision threshold of 20% is higher than the event rate of 5.6%, and **Δ**RU is defined as

$$\Delta\mathrm{RU} = (\Delta\mathrm{TP} - w \times \Delta\mathrm{FP})/\text{number of events},$$

with weight w for the odds of the decision threshold. In our example $\boldsymbol{\Delta}\text{RU} = (11 - 0.25 \times 5)/183 = 0.053$. **Δ**RU focuses on the improvement of using the prediction model to assign treatment over the baseline strategy with highest NB, either treat all or treat none (Baker, 2009). If the decision threshold is higher than the event rate, as in our example, this means that RU compares to treat none. Furthermore, the **Δ**RU divides the obtained improvement by the maximal improvement, which is the situation where treatment would be assigned to all individuals that develop the outcome of interest and to none without. The relation between **Δ**RU and **Δ**NB is $\boldsymbol{\Delta}\text{RU} = \boldsymbol{\Delta}\text{NB/event rate}$, in this situation of threshold > event rate (Van Calster et al., 2013). **Δ**NB considers the utility of the addition of a novel marker on an absolute scale, while **Δ**RU considers this utility relative to the maximum as defined by the event rate.

Finally, the weighted version of NRI only differs from **Δ**NB by a scaling factor: **Δ**NB is usually expressed in units of savings that result from a correct classification, whereas wNRI uses the savings in whatever unit is selected.

In our example, a 20% CHD risk decision threshold leads to **Δ**NB = 0.3%, which can be interpreted as that prediction with HDL cholesterol which increases the fraction of true positives identified in the population by 3 per 1000, without a change in false positives. **Δ**RU = 5.3%, which indicates a 5.3% relative gain in utility compared to the alternative of treating none. If we would assume that correct identification of a patient at risk of a CVD event results in savings of $100,000, whereas avoiding unnecessary treatment with statins saves $25,000, we obtain wNRI = $300 of average savings per person (0.3%*$100,000).

All decision analytic measures can be used to calculate the "test tradeoff", which indicates in how many persons the marker needs to be measured for a net increase in one true positive classification (i.e. identifying one additional person with the event as high risk who thereby qualifies for treatment) (Baker et al., 2012). This test tradeoff has similarities to the well-known concept of "number needed to treat" in trials and efficacy research. The test tradeoff is defined as 1/**Δ**NB. Hence, the test tradeoff for measuring HDL cholesterol was 335.

5.5 Sensitivity to choice of cut-off

In the weighted variants of summing sensitivity and specificity, the decision threshold is essential, which is defined by the harm to benefit ratio. As a sensitivity analysis, it is recommended to consider a range of possible thresholds. This can be displayed in a decision curve (Vickers and Elkin, 2006). We note that the model with HDL cholesterol has a small but consistently higher NB for most thresholds in the clinically most relevant range from 5% to 25% 10-year risks (Fig. 7).

6 Discussion

In this review, we considered various graphs and summary measures to assess the incremental value of a novel marker in predicting presence of disease (diagnosis) or the occurrence of an event over time (prognosis). We first examined continuous predictions, which is in line with the usual statistical approach to develop a model with and without the marker under study. Several informative graphs can be created (Figs. 1–5), with appealing summary measures such as the increase in explained variability and increase in AUC, or the continuous NRI. Second, we examined binary classifications, which can well be summarized in a reclassification table (Table 2). We propose an alternative design of such a table ("net reclassification table", Table 3) with a graph ("net reclassification risk graph", Fig. 6) to readily assess the incremental value of a marker for improving clinical risk classification. This presentation draws our attention to the number of reclassified persons, i.e. from low to high and from high to low risk. This focus was also central in Cook's early proposals for assessing reclassification (Cook et al., 2006; Cook, 2007). Second, our attention is drawn to the absolute risk estimates, which naturally should be higher for those reclassified as at high risk then for those reclassified as low risk. The difference

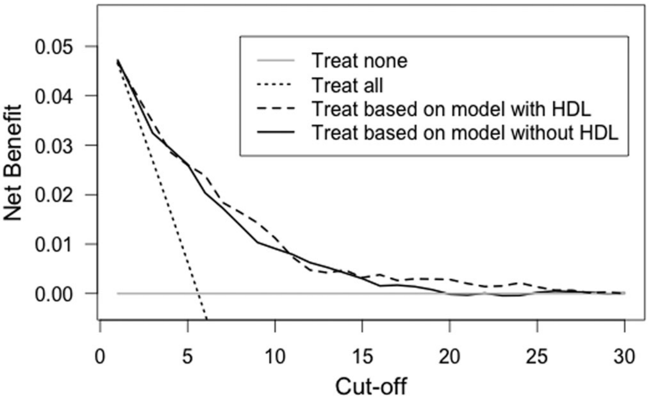

Figure 7 Decision curves comparing Net Benefit of four alternative strategies: treat all, treat none, and treatment decisions based on the predicted 10-year CHD risk (from models with or without HDL cholesterol) according to risk cut-offs between 0 and 30%.

in areas of the low→high and high→low bars indicates how many more subjects with events can be identified by the marker. This increase in sensitivity is a key attribute in many decision problems. The harms for false positive classifications, which relate to the loss of specificity, are typically less than the benefits for true positive classifications. We recognize that more research is necessary on the properties of the net reclassification risk graph and its potential applications. An obvious limitation is that the graph cannot readily consider multiple categories, since a three category classification would already lead to nine bars.

6.1 Interpretation of incremental value

A limitation of all our assessments is that it is difficult to define what increment in performance is "important", "substantial", or "meaningful". For statistical significance, the limit of 0.05 is widely accepted. A lower *p*-value is affected by the combination of effect size and sample size, and should hence not be sufficient to claim that a marker is important for better prediction. Effect size criteria are preferable, where a value of Cohen's D of 0.5 is widely accepted as reflecting a medium effect (Cohen, 1988). Some epidemiologists may also like estimates of relative risk per standard deviation increase in value of a marker. Pencina et al. (Pencina et al., 2012a) found that a medium effect size or OR of 1.65 per SD corresponds to a continuous NRI of 0.395. The increase in AUC depends on the AUC of the baseline model, e.g. Cohen's D of 0.5 corresponds to +0.05 for a baseline model with AUC of 0.65, whereas only to +0.02 for a baseline model with AUC of 0.80 (Pencina et al., 2012a). A third option is to consider cost-effectiveness criteria, where willingness to pay thresholds have been proposed in terms of euros (or dollars) per quality adjusted life-year (QALY) gained, for example $20,000 per QALY (Chapman et al., 2000; Postmus et al., 2012). Performing a full cost-effectiveness analysis may be too much effort for every marker studied for incremental value. An intermediate solution is to consider measures such as the **Δ**NB, **Δ**RU, wNRI, or a reciprocal summary measure, the test tradeoff. In our example, the test tradeoff for measuring HDL cholesterol was 1 in 335. Whether this number is valued relatively low or relatively high depends on the context. For example, if we consider starting statin treatment for those classified as high risk, and assume a relative risk of 0.73 for the reduction of CHD events (Taylor et al., 2013), we know that we can prevent 27% of the net reclassified CHD events. This absolute benefit should be weighted against the disadvantages of the marker measurement, such as financial costs, the burden or discomfort, and potential risks to the patient. The harms of overtreatment are already incorporated in the NB calculation, where the treatment threshold reflects the harm to benefit ratio (in our example 1:4, threshold 20% 10-year CHD risk).

6.2 Estimation issues

We focused on the apparent performance of models with and without markers. Especially in small data sets, it is well known that effect estimates for markers may be exaggerated. Such overfit leads to overoptimistic estimates of performance. Several internal validation techniques can be used to correct for optimism in performance, including crossvalidation, and bootstrap resampling (Steyerberg et al., 2001). Internally validated performance estimates can be derived for all measures as presented. Ideally, performance is determined on fully independent external validation data. Many differences may exist between the external validation data and the development setting, which makes that we may often be testing transportability of prediction models in time or place rather than validating marker performance in the more narrow sense of assessing reproducibility (Justice et al., 1999). It is imperative that the NRI for quantifying added value of a marker should only be calculated after recalibrating or refitting model predictions (Hilden and Gerds, 2014; Leening et al., 2014a). Such recalibration places the predicted risks with and without the marker at an equal and fair level with respect to calibration. The NRI then indicates the improvement in classification attributable to the marker, but conditional on adequate calibration. An alternative approach is to evaluate reclassification in the context of better clinical decision making. We may then take predicted risks from a model with and without a marker literally, i.e. without recalibration, and preferably use performance measures that are consistent with a decision making framework, such as the **Δ**NB, **Δ**RU, or wNRI (Van Calster et al., 2013).

A final limitation is that we focused on prediction of a binary endpoint, without considering the time-to-event nature of the data in our example. In several graphs and tables we conditioned on the event status, which is a simplification that is especially problematic if censoring occurs in many subjects before the end of follow-up. The net reclassification risk graph can readily be made with Kaplan-Meier estimates rather than observed event rates (Steyerberg and Pencina, 2010), and definitions for survival data are available for many performance measures, such as Nagelkerke's R^2, Harrell's or Uno's c statistic (equivalent to AUC for binary endpoints) (Uno et al., 2011), the IDI and NRI (Pencina et al., 2011), and the **Δ**NB (Vickers et al., 2008).

In conclusion, we recommend a clear distinction in the type of research question that is addressed in relation to the incremental value of a marker: better prediction or better classification. From a prediction perspective, continuous predictions may be considered with various graphs and summary measures. We are in favor of reclassification plots, but due to the lack of informativeness we recommend against publishing of ROC curves unless clinically motivated thresholds are indicated in such a graph. For summary measures, we are in favor of measures that indicate the explained variability (increase in Nagelkerke's or Pearson's R^2, the IDI), increase in discrimination (**Δ**AUC), or effect size (continuous NRI). If we move to the evaluation of a classification, the category-based NRI may be used, with several important caveats (Leening et al., 2014b). The Net Reclassification Risk graph may allow for a direct visual impression in the case of binary classification. We can then readily judge two important aspects of improved classification: which proportion of subjects is reclassified, and what is the difference in observed event rates between reclassified groups. When classification leads to decisions, decision-analytic summary measures such as **Δ**NB, **Δ**RU, and wNRI are adequate.

Acknowledgments Part of this research was supported by the Center for Translational Molecular Medicine (E.S. and M.V.: PCMM project; D.P.: TRIUMPH and PREDICCt projects). B.V.C. was supported by a postdoctoral fellowship from the Research Foundation–Flanders (FWO).

Conflict of interest

The authors have declared no conflict of interest.

References

Austin, P. C. and Steyerberg, E. W. (2013). Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. *Statistics in Medicine* **32**, 661–672.

Baker, S. G. (2009). Putting risk prediction in perspective: relative utility curves. *Journal of National Cancer Institute* **101**, 1538–1542.

Baker, S. G., Cook, N. R., Vickers, A., and Kramer, B. S. (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society Series A* **172**, 729–748.

Baker, S. G., Van Calster, B., and Steyerberg, E. W. (2012). Evaluating a new marker for risk prediction using the test tradeoff: an update. *The International Journal of Biostatistics* **8**, 22499728.

Chapman, R. H., Stone, P. W., Sandberg, E. A., Bell, C., and Neumann, P. J. (2000). A comprehensive league table of cost-utility ratios and a sub-table of "panel-worthy" studies. *Medical Decision Making* **20**, 451–467.

Cleveland, W. S. (1985). *The elements of graphing data*. Wadsworth Advanced Books and Software: Monterey, CA.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, New Jersey, NJ.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928–935.

Cook, N. R., Buring, J. E., and Ridker, P. M. (2006). The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine* **145**, 21–29.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, New York, NY.

Hilden, J. and Gerds, T. A. (2014). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine*, in press.

Justice, A. C., Covinsky, K. E., and Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine* **130**, 515–524.

Leening, M. J. and Steyerberg, E. W. (2013). Fibrosis and mortality in patients with dilated cardiomyopathy. *JAMA* **309**, 2547–2548.

Leening, M. J., Steyerberg, E. W., Van Calster, B., D'Agostino, R. B., and Pencina, M. J. (2014a). Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. *Statistics in Medicine*, in press.

Leening, M. J., Vedder, M. M., Witteman, J. C., Pencina, M. J., and Steyerberg, E. W. (2014b). Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Annals of Internal Medicine* **160**, 122–131.

Localio, A. R. and Goodman, S. (2012). Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Annals of Internal Medicine* **157**, 294–295.

McGeechan, K., Macaskill, P., Irwig, L., Liew, G., and Wong, T. Y. (2008). Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. *Archives of Internal Medicine* **168**, 2304–2310.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692.

Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Jornal of Medicine* **302**, 1109–1117.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science* **4**, 453–454.

Pencina, M. J., D'Agostino, R. B., Pencina, K. M., Janssens, A. C., and Greenland, P. (2012a). Interpreting incremental value of markers added to risk prediction models. *American Journal of Epidemiology* **176**, 473–481.

Pencina, M. J., D'Agostino, R. B., Sr., D'Agostino, R. B., Jr., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172; discussion 207–212.

Pencina, M. J., D'Agostino, R. B., Sr., and Demler, O. V. (2012b). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in Medicine* **31**, 101–113.

Pencina, M. J., D'Agostino, R. B., Sr., and Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* **30**, 11–21.

Pencina, M. J., D'Agostino, R. B., and Vasan, R. S. (2010). Statistical methods for assessment of added usefulness of new biomarkers. *Clinical Chemistry and Laboratory Medicine* **48**, 1703–1711.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., and Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* **167**, 362–368.

Postmus, D., de Graaf, G., Hillege, H. L., Steyerberg, E. W., and Buskens, E. (2012). A method for the early health technology assessment of novel biomarker measurement in primary prevention programs. *Statistics in Medicine* **31**, 2733–2744.

Ridker, P. M., Buring, J. E., Rifai, N., and Cook, N. R. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *JAMA* **297**, 611–619.

Royston, P., and Altman, D. G. (2010). Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine* **29**, 2508–2520.

Steyerberg, E. W. (2009). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer, New York, NY.

Steyerberg, E. W., Harrell, F. E., Jr., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., and Habbema, J. D. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* **54**, 774–781.

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., and Group, P. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine* **10**, e1001381.

Steyerberg, E. W. and Pencina, M. J. (2010). Reclassification calculations for persons with incomplete follow-up. *Annals of Internal Medicine* **152**, 195–197.

Taylor, F., Huffman, M. D., Macedo, A. F., Moore, T. H., Burke, M., Davey Smith, G., Ward, K., and Ebrahim, S. (2013). Statins for the primary prevention of cardiovascular disease. *The Cochrane Database of Systematic Reviews*, **1**, CD004816.

Tjur, T. (2009). Coefficients of determination in logistic regression models—a new proposal: the coefficient of discrimination. *The American Statistician* **63**, 366–372.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**, 1105–1117.

Van Calster, B., Steyerberg, E.W., D'Agostino, R. B., Sr., and Pencina, M. J. (2014). Sensitivity and specificity can change in opposite directions when new predictive markers are added to risk models. *Medical Decision Making* **34**, 513–522.

Van Calster, B., Vickers, A. J., Pencina, M. J., Baker, S. G., Timmerman, D., and Steyerberg, E. W. (2013). Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. *Medical Decision Making* **33**, 490–501.

Vickers, A. J., Cronin, A. M., Elkin, E. B., and Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making* **8**, 53.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **26**, 565–574.