American Journal of Epidemiology
Vol. 176, No. 6
DOI: 10.1093/aje/kws207
Advance Access publication:
August 8, 2012

Practice of Epidemiology

Interpreting Incremental Value of Markers Added to Risk Prediction Models

Michael J. Pencina*, Ralph B. D'Agostino, Karol M. Pencina, A. Cecile J. W. Janssens, and Philip Greenland

* Correspondence to Dr. Michael J. Pencina, Department of Biostatistics, Boston Unviersity, Framingham Heart Study, Harvard Clinical Research Institute, Room 328, CrossTown, 3rd Floor, 801 Massachusetts Avenue, Boston, MA 02118 (e-mail: mpencina@bu.edu).

Initially submitted June 20, 2011; accepted for publication October 6, 2011.

The discrimination of a risk prediction model measures that model's ability to distinguish between subjects with and without events. The area under the receiver operating characteristic curve (AUC) is a popular measure of discrimination. However, the AUC has recently been criticized for its insensitivity in model comparisons in which the baseline model has performed well. Thus, 2 other measures have been proposed to capture improvement in discrimination for nested models: the integrated discrimination improvement and the continuous net reclassification improvement. In the present study, the authors use mathematical relations and numerical simulations to quantify the improvement in discrimination offered by candidate markers of different strengths as measured by their effect sizes. They demonstrate that the increase in the AUC depends on the strength of the baseline model, which is true to a lesser degree for the integrated discrimination improvement. On the other hand, the continuous net reclassification improvement depends only on the effect size of the candidate variable and its correlation with other predictors. These measures are illustrated using the Framingham model for incident atrial fibrillation. The authors conclude that the increase in the AUC, integrated discrimination improvement, and net reclassification improvement offer complementary information and thus recommend reporting all 3 alongside measures characterizing the performance of the final model.

area under curve; biomarkers; discrimination; risk assessment; risk factors

Abbreviations: ΔAUC, change in area under the receiver operating characteristic curve; AUC, area under the receiver operating characteristic curve; BNP, B-type natriuretic peptide; CRP, C-reactive protein; IDI, integrated discrimination improvement; NRI, net reclassification improvement; NRI(>0), continuous net reclassification improvement.

Editor's note: *Invited commentaries on this article appear on pages 482 and 488, and the authors' response appears on page 492.*

Risk prediction models have been successfully developed in all major fields of modern medicine, including cardiovascular disease, cancer, and diabetes (1–7). The performance of said models is assessed by both their calibration and discrimination (8). Calibration addresses the question of how closely the model-based risk estimates align with the observed outcomes. Discrimination focuses on a model's ability to distinguish between subjects who will (or did) develop the event of interest from those who will (did) not. "Good" prediction models can then be developed into risk prediction rules. Such rules may classify people into 2 (e.g., high vs. low risk; treat vs. do not treat) or 3 (e.g., high, intermediate, or low risk; treat pharmacologically, introduce lifestyle intervention, or do not act) medical decision categories based on preselected thresholds (9).

The importance and usefulness of risk prediction models has been recognized by the medical community, and in some cases, their use has even been incorporated into clinical treatment guidelines (9). However, as new risk factors

or markers are discovered (10–12), it has become imperative to identify the ones that merit routine measurement and incorporation into the aforementioned prediction models and rules (13–16). Recently, an expert panel in the cardiovascular field presented the phases of an evaluation of novel risk markers (14). Briefly, after the marker has been shown to predict disease onset, it must demonstrate that it adds incremental value to risk prediction models that contain standard factors; additionally, it needs to show clinical utility (it must change the predicted risk enough to warrant a change in therapy), and it needs to be cost-effective. In the present study, we investigated the incremental value that the new markers add to prediction models with standard factors. This is a more basic and more purely statistical assessment that ignores the costs and utilities while focusing on measures that do not lose information because of reliance on risk categories. As recommended by Hlatky et al. (14), this assessment should be followed by an evaluation of clinical utility and a cost-effectiveness analysis.

Because new candidate markers must be associated with the onset of disease after controlling for risk factors already included in the baseline model, it is necessary to assess their incremental value for risk prediction models in terms other than statistical significance. Furthermore, the increasing availability of large databases will lead to more markers crossing the threshold of statistical significance. Here, our goals were 2-fold. First, we intended to assess the impact on model discrimination exerted by the addition of candidate markers of various strengths. For continuous markers, we defined their strength in terms of effect size and considered the following 3 measures of improvement in model performance: the change in the area under the curve (ΔAUC), the integrated discrimination improvement (IDI), and the net reclassification improvement (NRI) (17–19) (Web Appendix 1, available at http://aje.oxfordjournals.org/). Second, we wanted to determine how these 3 measures were affected by differences in the strength of the baseline model based on standard risk factors. We accomplished this using numerical simulations and exploiting the existing links between these 3 measures and the concept of effect size in the context of normally distributed variables, as shown by Pencina et al. (20), Demler et al. (21), and Royston and Altman (22).

CONCEPT OF EFFECT SIZE

Let X be a normally distributed predictor with means among subjects with and without events denoted by μ_1 and μ_0, respectively. Furthermore, assume a common standard deviation among these 2 groups and denote it with σ. A measure of effect size introduced by Cohen (23) is defined as $d = (\mu_1 - \mu_0)/\sigma$. Its sample estimator, which uses the pooled variance formula, is known as Hedges g (24). For practical applications, Cohen proposed ad hoc labels quantifying the strength of the effect size resulting from the above equation. An effect size of $d = 0.8$ is considered strong, $d = 0.2$ is considered weak, and $d = 0.5$ is considered medium. Cohen himself admitted that these numbers are arbitrary and suggested that they be used only if there is no other way to deduce the importance of the observed effect size.

Because the effect size is not commonly used in risk prediction, to determine whether Cohen's benchmarks can be viewed as reasonable in this setting, we needed to translate them into the more familiar metrics. We did this under the assumption of normality, where such translation is possible (22, 25). We used measures commonly provided in publications focused on risk prediction, including the odds ratio per 1 standard deviation as a measure of association, the sensitivity at the point where the specificity equals 0.85 (motivated by the approximate specificity of the Framingham risk prediction rule established by D'Agostino et al. (1)) as a measure of diagnostic accuracy, and the area under the receiver operating characteristic curve (AUC) and discrimination slope for different event rates as measures of model discrimination. In Table 1, we show how the effect sizes of 0.2, 0.5, and 0.8 translate into the above measures. The mathematical derivations are summarized in Web Appendix 2.

We observed that a variable with a "weak" effect size of 0.2 translated into a weak association, diagnostic performance, and model discrimination. On the other hand, a variable with a "strong" effect size of 0.8 led to a model with a reasonable AUC of 0.71. On the basis of the direct interpretation of the performance metrics, we may not consider the AUC of 0.71 and sensitivity of 0.41 strong. However, these results were strikingly similar to the effect of age as a single predictor in 10-year cardiovascular risk model based on the Framingham data (1). Because age is known to be

Table 1. Relation Between Different Measures of Effect Size for Normal Data

Cohen's *d*	Odds Ratio[a]	Sensitivity[b]	AUC	Discrimination Slope By Varying Percentages of Event Rates			
				5%	10%	20%	50%
0.2	1.22	0.20	0.56	0.002	0.004	0.006	0.010
0.5	1.65	0.30	0.64	0.013	0.024	0.040	0.059
0.8	2.23	0.41	0.71	0.037	0.064	0.101	0.139

Abbreviation: AUC, area under the receiver operating characteristic curve.
[a] Per 1 standard deviation.
[b] Specificity = 0.85.

the best discriminator in cardiovascular risk prediction (provided its distribution is wide enough in the population of interest), based on comparative interpretation, we can still argue for the "strong" label for effect sizes above 0.8. The effect size of 0.5 falls directly in the middle, justifying its label.

The above discussion focused on the univariate case, in which the strength of one variable translates into the strength of the model. It provided a simple context through which we could derive the notions of effect magnitude for a variable and the corresponding model. In what follows, we focused on an arguably more pertinent issue: quantifying the impact of variables added to a risk score that already contains a set of standard factors. We addressed the question of the extent to which the strength of a variable translates into improvement in model discrimination and rely on comparative interpretation to derive heuristic benchmarks for small, medium, and large incremental values.

STRENGTH OF NEW PREDICTORS AND IMPROVEMENT IN DISCRIMINATION

Improvement in the discrimination of risk prediction models can be quantified in numerous ways (13, 15, 18, 19). A natural approach takes the difference in discrimination metrics between the models with and without the new predictor. The ΔAUC is produced in this manner and so is the IDI, defined as a difference in discrimination slopes (19). The relative IDI (26) can be calculated as the ratio of IDI over the discrimination slope of the model without the new predictor ("baseline model"). A different metric, called the continuous NRI (NRI(>0)), is obtained when we focus on the relative increase in the predicted probabilities for subjects who experience events and the decrease for subjects who do not (19) (Web Appendix 1).

We were interested in determining the magnitude of improvement rather than testing the hypothesis that said improvement was greater than zero. We therefore determined the approximate improvements in discrimination that were incurred when adding weak, medium, and strong new predictors to any baseline model. Because a vast majority of the risk prediction algorithms are based on generalized linear models of some form, we assumed that the predicted probabilities for the event from the baseline model were uniquely determined by its linear predictor. Furthermore, it is not unreasonable to assume that this linear predictor was distributed normally. For technical reasons, we assumed its normal distribution within the event categories with equal covariance matrices.

First, we considered a novel marker, also distributed normally, with equal covariance matrices within event groups. Pencina et al. (20) and Demler et al. (21) have shown that in this case, the ΔAUC, the IDI, and the NRI(>0) are functions of a generalized measure of separation known as the squared Mahalanobis distance, or M^2 (27). In the case of the IDI, we also needed to know the ratio of the nonevents to events prevalence (or incidence). Exact formulas are given in Web Appendix 3. If the predictors were uncorrelated, the M^2 would reduce to the sum of squared effect sizes. However, because the measures of interest depended only on M^2 without any loss of generality, we could assume that the predictors were not correlated. Indeed, for any new predictor correlated with predictors already in the model, we can find a predictor that is uncorrelated and results in the same increase in the M^2. Hence, in the following text, the additional predictor will be conditionally (within event categories) uncorrelated with the rest.

Because the magnitude of the AUC is the most well understood of the measures, we focused on baseline models with AUCs ranging from 0.50 (useless) to 0.90 (excellent), in increments of 0.05. Using the associations given in Web Appendix 3, we translated these AUCs into M^2s for the baseline models. Adding an uncorrelated normal predictor with an effect size d increased the M^2 by d^2. Said increases were then transformed into measures of improvement in discrimination using the identities given in Web Appendix 3. The IDI was calculated for event rates equal to 0.05, 0.10, 0.20, and 0.50, representing a wide range of possible scenarios. The results are presented in Table 2.

A few observations merit particular attention. First, we noticed that consistent with empirical evidence (28), improvement in the AUC depended strongly on the baseline model. If we started with a poor model that had an AUC of 0.60, a new predictor with a strong effect of 0.8 could raise the AUC by 0.13. The same predictor added to a very good baseline model with an AUC of 0.85 would raise that AUC by merely 0.03. The same was true for both medium and weak predictors, with their contribution becoming more miniscule as a function of the baseline model's AUC. By contrast, this phenomenon was much weaker for the IDI. In particular, the IDI was stable as a function of the baseline model for an event rate of 10%. For lower event rates (e.g., 5%), the IDI increased as a function of the strength of the baseline model, and for larger rates (i.e., 20% and 50%), it decreased. The weaker effect in the IDI compared with that in the ΔAUC can be partially explained by the discrimination slopes, particularly those for smaller event rates, which are much further from their maximum of 1.00 than are the AUCs. Hence, there remains more room for improvement.

As suggested by formula 3 in Web Appendix 3, the NRI (>0) depends mainly on the effect size of the added predictor rather than on the strength of the baseline model. This was confirmed in Table 2, where the NRI(>0) values were constant across the baseline models: 0.62 for a strong predictor, 0.39 for a medium predictor, and 0.16 for a weak, uncorrelated predictor added to the baseline model. This illustrates an important property of the NRI(>0): Its value depends only on the strength of its association with the outcome and not on the strength of the baseline model. Of note, our assumption of uncorrelatedness simplified things here; the NRI(>0) still captured the impact of correlation and penalized those markers that might be strongly associated with the outcome but also correlated with variables already in the model.

Further, in Figure 1, we plotted the discrimination slope as a function of the M^2 under the assumption of normality. We noticed that regardless of the event rate, for values of the M^2 below 3, the relation was not far from a linear relation with no intercept. Assuming this simple approximation is reasonable, we have a slope $\approx \beta_1 \times M^2$. If we also assume the baseline model is comprised of p uncorrelated

Table 2. Improvement in Discrimination as a Function of Discrimination of Baseline With Single Normal Predictor and Normal New Predictor: Theoretical Calculations

Model and Metric	Value[a]								
Baseline									
AUC	0.500	0.550	0.600	0.650	0.700	0.750	0.800	0.850	0.900
M	0.000	0.178	0.358	0.545	0.742	0.954	1.190	1.466	1.812
M^2	0.000	0.032	0.128	0.297	0.550	0.910	1.417	2.148	3.285
Slope_5^b	0.000	0.002	0.006	0.016	0.031	0.057	0.097	0.161	0.264
Slope_{10}	0.000	0.003	0.012	0.028	0.055	0.094	0.151	0.231	0.346
Slope_{20}	0.000	0.005	0.021	0.047	0.087	0.142	0.213	0.306	0.426
Slope_{50}	0.000	0.008	0.031	0.069	0.122	0.189	0.270	0.369	0.489
Baseline plus uncorrelated normal predictor with large effect size of 0.8									
ΔAUC	0.214	0.169	0.132	0.103	0.080	0.061	0.045	0.031	0.019
ΔM	0.800	0.642	0.518	0.423	0.349	0.291	0.244	0.204	0.169
ΔM^2	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640	0.640
IDI_5^b	0.037	0.038	0.040	0.043	0.047	0.052	0.056	0.058	0.056
IDI_{10}	0.064	0.065	0.066	0.068	0.071	0.072	0.071	0.067	0.058
IDI_{20}	0.101	0.101	0.100	0.098	0.095	0.089	0.082	0.071	0.058
IDI_{50}	0.139	0.138	0.132	0.124	0.113	0.101	0.087	0.072	0.055
NRI(>0)	0.622	0.622	0.622	0.622	0.622	0.622	0.622	0.622	0.622
Baseline plus uncorrelated normal predictor with medium effect size of 0.5									
ΔAUC	0.138	0.096	0.068	0.049	0.036	0.027	0.019	0.013	0.008
ΔM	0.500	0.353	0.257	0.195	0.153	0.123	0.101	0.083	0.068
ΔM^2	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
IDI_5^b	0.013	0.013	0.014	0.015	0.017	0.019	0.021	0.023	0.022
IDI_{10}	0.024	0.024	0.025	0.026	0.027	0.028	0.028	0.027	0.023
IDI_{20}	0.040	0.040	0.040	0.039	0.038	0.036	0.033	0.029	0.023
IDI_{50}	0.059	0.058	0.056	0.052	0.047	0.042	0.036	0.029	0.023
NRI(>0)	0.395	0.395	0.395	0.395	0.395	0.395	0.395	0.395	0.395
Baseline plus uncorrelated normal predictor with small effect size of 0.2									
ΔAUC	0.056	0.025	0.014	0.009	0.006	0.005	0.003	0.002	0.001
ΔM	0.200	0.090	0.052	0.036	0.026	0.021	0.017	0.014	0.011
ΔM^2	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040
IDI_5^b	0.002	0.002	0.002	0.002	0.003	0.003	0.003	0.004	0.004
IDI_{10}	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
IDI_{20}	0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.004
IDI_{50}	0.010	0.010	0.009	0.009	0.008	0.007	0.006	0.005	0.004
NRI(>0)	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.160

Abbreviations: AUC, area under the receiver operating characteristic curve; IDI, integrated discrimination improvement; NRI, net reclassification improvement.

[a] Values show changes after adding a new variable as a function of the corresponding baseline metric.

[b] Numerical subscript indicates percent of events, that is, slope_5 means discrimination slope when event rate is 5%.

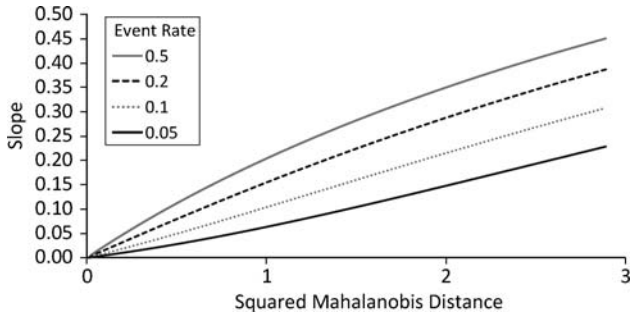

Figure 1. Discrimination slope as function of squared Mahalanobis distance.

predictors, each with an equal effect size of d, and the new model adds another uncorrelated predictor with an effect size d, we get the following: slope (baseline) $\approx \beta_1 \times p \times d^2$ and slope (new) $\approx \beta_1 \times (p+1) \times d^2$. Hence, the IDI $\approx \beta_1 \times d^2$ and the relative IDI $\approx 1/p$. This introduces a useful benchmark for the comparative interpretation of the relative IDI in reference to variables already included in the model: If the new predictor has a squared effect size similar to the average of the squared effect sizes of the p variables included in the baseline model, then the relative IDI should be $1/p$. This offers a simple comparative benchmark – the relative IDI over and above the inverse of the number of variables in the baseline model indicates a predictor better than the average of the rest, whereas a value below the inverse suggests the new predictor is worse than that average. We note that the criterion of proportionality to the square of the effect size was justified by the fact that the squared Mahalanobis distances or squared effect sizes are additive for uncorrelated markers, whereas the effect sizes themselves are not.

The results presented thus far were derived under specific and restrictive assumptions. However, even though most variables are not normally distributed, the majority of continuous predictors can be appropriately transformed to be not far from normal. As mentioned earlier, the assumption of uncorrelatedness is not restrictive, and formulas analogous to those given in Web Appendix 3 exist in cases of unequal covariance matrices. Separate investigation is necessary to extend the results to conditions with binary or categorical predictors. Because any nonordinal categorical predictor can be defined by a set of "dummy" variables, we focused on binary variables.

To obtain results parallel to those presented in Table 2, we resorted to numerical simulations. We simulated 2 independent variables from multivariate normal distributions within the event categories, with the first representing the linear predictor of the baseline model and the second dichotomized at a prespecified threshold to represent a new added binary predictor. Effect sizes for the linear predictors matched those listed in the second row of Table 2 and were set for the second dichotomized predictor at 0.2, 0.5, and 0.8. Because the strength of dichotomous predictors depends on the prevalence of the "exposed" among the events and nonevents, we considered various dichotomization thresholds. For brevity, we present only the results for a threshold that guarantees a specificity of 0.85 if the variable were used as a single predictor. Sample size was set at 1,000,000 to obtain results on a population level and the event rates matched those presented in Table 2. Two logistic regressions models were fitted, one with the linear predictor only and second with the linear predictor and the additional binary variable. The quantities of interest (same as in Table 2) were calculated using published methods (17–19, 29). Of note, changes in the Mahalanobis distance were obtained by inverting the AUC, as it is not defined in the nonnormal case. The results are given in Table 3.

We observed that the contribution of the binary risk factor was smaller than that of the corresponding continuous predictor, which underscores the inefficiency of dichotomization. Not surprisingly, the impact of the binary risk factor depended on the prevalence of the "exposed." Table 3 represents only one scenario; in several other cases, the impact of the binary risk factor varied with this prevalence but never exceeded the impact of the corresponding continuous predictor (results not shown). The maximum NRI(>0) was reached when the Youden's index (30) of the binary risk factor was maximized and it approached the value obtained for the continuous predictor. Furthermore, we observed that the general patterns seen in Table 2 still held: The NRI(>0) remained constant regardless of the strength of the baseline model, which was not true for the ΔAUC or the IDI, although the IDI was much less affected.

PRACTICAL EXAMPLE

We illustrate the concepts described in this article with an example from the Framingham Heart Study. The focus was on the assessment of 10-year risk of atrial fibrillation in people free of the condition at baseline between 1995 and 1998. A sample of 3,120 Framingham participants aged 29 to 86 years were available for analysis. A total of 203 cases of atrial fibrillation occurred within 10 years of follow-up. These data have been analyzed previously by Schnabel et al. (31). Three logistic regression models were fit. The first included sex and baseline age, as well as standard risk factors; the second included all of the above variables plus the natural logarithm of B-type natriuretic peptide (BNP); and the third included the natural logarithm of C-reactive protein (CRP) instead of BNP. The AUC was estimated using the c statistic (17, 29), and the Mahalanobis distance was calculated by inverting the AUC using the relation presented in Web Appendix 2. The discrimination slopes, IDI, relative IDI, and NRI(>0) were computed using the methods of Pencina et al. (18, 19). The results are presented in Table 4.

Both biomarkers were significantly associated with the outcome after controlling for other risk factors (per 1 standard deviation in log-biomarker, odds ratio = 1.68 for BNP and odds ratio = 1.24 for CRP). Log-transformed BNP had a much stronger impact on model performance than did CRP. BNP increased the AUC from 0.774 to 0.805, whereas CRP raised it only to 0.780. On the absolute scale,

Table 3. Improvement in Discrimination as a Function of Discrimination of Baseline With a Single Normal Predictor and a Binary New Predictor: Estimated Results

Model and Metric	Value[a]								
Baseline									
AUC	0.500	0.550	0.600	0.650	0.700	0.750	0.800	0.850	0.900
M	0.004	0.181	0.361	0.548	0.746	0.956	1.191	1.463	1.779
M^2	0.000	0.033	0.130	0.301	0.556	0.915	1.418	2.142	3.165
Slope_5^b	0.000	0.002	0.006	0.016	0.016	0.057	0.097	0.162	0.265
Slope_{10}	0.000	0.003	0.012	0.028	0.028	0.094	0.150	0.231	0.346
Slope_{20}	0.000	0.005	0.020	0.047	0.047	0.141	0.213	0.305	0.425
Slope_{50}	0.000	0.008	0.031	0.069	0.069	0.188	0.269	0.368	0.488
Baseline plus uncorrelated normal predictor with large effect size of 0.8									
ΔAUC	0.129	0.108	0.086	0.066	0.050	0.037	0.027	0.017	0.014
ΔM	0.468	0.397	0.327	0.264	0.211	0.172	0.140	0.108	0.113
ΔM^2	0.223	0.301	0.344	0.358	0.360	0.358	0.354	0.327	0.416
IDI_5^b	0.024	0.024	0.026	0.028	0.031	0.034	0.036	0.037	0.036
IDI_{10}	0.041	0.042	0.043	0.044	0.046	0.046	0.045	0.042	0.036
IDI_{20}	0.066	0.066	0.065	0.063	0.060	0.056	0.050	0.043	0.034
IDI_{50}	0.083	0.082	0.078	0.073	0.067	0.059	0.051	0.042	0.032
NRI(>0)	0.523	0.523	0.523	0.523	0.523	0.523	0.523	0.523	0.523
Baseline plus uncorrelated normal predictor with medium effect size of 0.5									
ΔAUC	0.073	0.055	0.039	0.027	0.020	0.014	0.010	0.007	0.007
ΔM	0.260	0.199	0.145	0.106	0.081	0.065	0.052	0.040	0.052
ΔM^2	0.070	0.112	0.125	0.128	0.127	0.128	0.127	0.119	0.188
IDI_5^b	0.008	0.008	0.009	0.009	0.010	0.012	0.013	0.013	0.013
IDI_{10}	0.014	0.014	0.015	0.015	0.016	0.016	0.016	0.015	0.013
IDI_{20}	0.023	0.023	0.023	0.023	0.022	0.020	0.018	0.016	0.013
IDI_{50}	0.031	0.031	0.030	0.028	0.025	0.022	0.019	0.016	0.012
NRI(>0)	0.296	0.296	0.296	0.296	0.296	0.296	0.296	0.296	0.296
Baseline plus uncorrelated normal predictor with small effect size of 0.2									
ΔAUC	0.025	0.013	0.007	0.004	0.003	0.002	0.002	0.001	0.001
ΔM	0.088	0.045	0.025	0.017	0.012	0.010	0.008	0.006	0.008
ΔM^2	0.009	0.018	0.019	0.019	0.019	0.020	0.020	0.017	0.029
IDI_5^b	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	0.002
IDI_{10}	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
IDI_{20}	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002
IDI_{50}	0.005	0.005	0.005	0.004	0.004	0.003	0.003	0.002	0.002
NRI(>0)	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104

Abbreviations: AUC, area under the receiver operating characteristic curve; IDI, integrated discrimination improvement; NRI, net reclassification improvement.

[a] Values show changes after adding a new variable as a function of the corresponding baseline metric.

[b] Numerical subscript indicates percent of events, that is, slope_5 means discrimination slope when event rate is 5%.

Table 4. Increase in Discrimination of Atrial Fibrillation Risk Model After Addition of B-Type Natriuretic Peptide and C-Reactive Protein

Model and Metric	Value
Model with standard risk factors[a] and baseline age	
AUC	0.774
M^b	1.065
M^{2b}	1.134
Slope	0.078
Contribution of BNP[c] to standard risk factors and baseline age model	
ΔAUC	0.031
ΔM^b	0.153
ΔM^{2b}	0.349
IDI	0.027
Relative IDI	0.352
NRI(>0)	0.474
Contribution of CRP[c] to standard risk factors and baseline age model	
ΔAUC	0.006
ΔM^b	0.026
ΔM^{2b}	0.056
IDI	0.003
Relative IDI	0.040
NRI(>0)	0.237

Abbreviations: AUC, area under the receiver operating characteristic curve; BNP, B-type natriuretic peptide; CRP, C-reactive protein; IDI, integrated discrimination improvement; NRI, net reclassification improvement.

[a] Sex, body mass index, systolic blood pressure, electrocardiogram PR interval, hypertension treatment, heart valve disease, and heart failure.

[b] Based on inverting the AUC.

[c] Natural logarithmically transformed.

the difference in mean predicted probabilities of event between subjects with and without event was 0.078 for the baseline model. It increased to 0.105 when BNP was added and to 0.081 when CRP was added. These increases translate to relative IDIs of 0.352 and 0.038 for BNP and CRP, respectively. The first value was well above and the second value was well below the average contribution of the other 8 variables already included in the model, which equaled 0.125. Finally, the NRI(>0) was 0.474 for BNP and 0.237 for CRP, corresponding to an uncorrelated continuous predictors with more than a medium and weak effect sizes, respectively.

On the basis of these results, we concluded that BNP is more promising in terms of improving predictive value than is CRP. All metrics considered paint a consistent picture that can be expected in cases of continuous unimodal predictors and baseline models that are not exceptionally strong. Furthermore, the facts that the strength of model improvement achieved by the inclusion of BNP corresponded to the effect expected when a new uncorrelated predictor of more than a medium effect size was added and the relative impact of BNP exceeds that of the average predictor already in the model suggest that BNP should be considered for further evaluation in terms of clinical utility and cost-effectiveness.

CLINICAL IMPLICATIONS

The AUC provides a familiar summary for the discriminatory ability of risk models. Its increment, ΔAUC, does not have a direct interpretation of its own beyond what it means: a difference between the AUCs of models with and without the new predictor(s). This increment should always be reported together with the AUC of the baseline model to put it in the proper context. New predictors of different strengths are needed to achieve the same ΔAUC depending on the strength of the baseline model, which could lead to the opposite conclusions about the same candidate marker evaluated in studies with baseline models of different strengths. For example, to increase the AUC from 0.50 to 0.55, a new predictor with a weak effect size below 0.2 will suffice; to increase the AUC from 0.80 to 0.85, we need a new predictor with a strong effect size (above 0.8). These increases can be translated into changes in sensitivity induced while holding specificity fixed; for example, when specificity is set at 0.85, an increase in AUC from 0.50 to 0.55 implies an increase in sensitivity from 0.15 to 0.20, and an increase in AUC from 0.80 to 0.85 implies an increase in sensitivity from 0.56 to 0.66. The ΔAUC is the preferred metric in settings in which the focus is on the model itself rather than on the variables that are to be added.

On the other hand, when assessing the true discriminatory potential of a new predictor in contrast to other predictors, especially from different or nonhomogeneous studies, the NRI(>0) is probably the best metric. It captures the marginal strength of the new predictor after accounting for correlations with variables included in the baseline model, and it can be used to compare predictors with different statistical distributions, a desirable feature not available when using the odds ratios, hazard ratios, or effect sizes. Our analysis suggested simple interpretation benchmarks for the NRI(>0) based on the effect size labels proposed by Cohen: NRI(>0) values above 0.6 should be considered strong, those around 0.4 should be considered intermediate, and those below 0.2 should be considered weak. Furthermore, NRI(>0) can be viewed as a limiting case of the category-based NRI (18), in which each unique predicted probability forms its own category. This gives the NRI(>0) an interpretation as a summary measure quantifying the correct upward versus downward movement in model-based predicted probabilities for events and nonevents.

The 2 different perspectives offered by the ΔAUC and NRI(>0) are bridged by the IDI. The IDI is not as easily influenced by the strength of the baseline model as ΔAUC, but at the same time, it has a direct connection to a model performance metric: the discrimination slope. Moreover, its magnitude has a direct interpretation as the amount by which we increased the separation of mean predicted probabilities for events and nonevents. In our example, BNP had an IDI of 0.027, having increased the separation of mean predicted

probabilities for events and nonevents from 0.078 to 0.105. The magnitude of this improvement can be presented in a comparative context based on the simple benchmark derived for the relative IDI. If the predictor is of similar strength to those predictors already present in the model, the relative IDI should be equal to the inverse of the number of predictors.

The IDI and discrimination slope operate on the absolute scale of model-based predicted probabilities, which is particularly desirable when we are concerned with absolute risks. However, this also makes it dependent on the overall event rate, and thus discrimination slopes and IDIs cannot be compared among studies with different rates and can be heavily influenced by model calibration. This dependence may also complicate interpretation, as the meaning of the IDI is different in studies with different event rates. That is why it can be useful to look at the relative IDI, which standardizes the observed increment to the discrimination slope of the baseline model.

CONCLUSIONS

In the present study, we focused on statistical measures of incremental value of new predictors. We assessed and quantified the relation between the strength of the predictor in terms of its effect size and its incremental value when added to a risk prediction model and provided simple guidelines for comparative interpretation. Because the ΔAUC, IDI, and NRI(>0) offer complementary information, in many settings it will be important to report all 3 alongside measures characterizing the performance of the final model to provide a complete assessment of incremental predictive value. Further research is needed to determine if the relations presented here hold under less restrictive assumptions.

The measures described here form the "first line" of assessment that can be particularly helpful in prescreening markers with limited promise. On the other hand, if a given predictor is deemed worthy of future exploration, a formal cost-benefit analysis should be undertaken. The weighted NRI with categories (19) or the net benefit analysis (32, 33) are promising simple options for the second step, directed more towards establishing clinical utility.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Boston University, Boston, Massachusetts (Michael J. Pencina); Department of Mathematics and Statistics, Boston University, Boston, Massachusetts (Michael J. Pencina, Ralph B. D'Agostino, Karol M. Pencina); Harvard Clinical Research Institute, Boston, Massachusetts (Michael J. Pencina, Ralph B. D'Agostino); Department of Epidemiology, Erasmus University Medical Center, Rotterdam, the Netherlands (A. Cecile J. W. Janssens); and Feinberg School of Medicine, Northwestern University, Chicago Illinois (Philip Greenland).

This work was supported by the National Institutes of Health/American Recovery and Reinvestment Act Risk Prediction of Atrial Fibrillation (grant 1 RC1HL101056; Michael J. Pencina and Karol M. Pencina); the National Heart, Lung, and Blood Institute's Framingham Heart Study (contract N01-HC-25195; Michael J. Pencina and Ralph B. D'Agostino); the Center for Medical Systems Biology in the framework of the Netherlands Genomics Initiative and the Netherlands Organisation for Scientific Research (A. Cecile J. W. Janssens); and the Northwestern University Clinical and Translational Sciences Institute (grant UL1RR025741; Philip Greenland).

Conflict of interest: none declared.

REFERENCES

1. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care. *Circulation*. 2008;117(6):743–753.
2. Wilson PWF, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97(18):1837–1847.
3. D'Agostino RB, Wolf PA, Belanger A, et al. Stroke risk profile: adjustment for antihypertensive medication. *Stroke*. 1994;25(1):40–43.
4. Schnabel RB, Sullivan LM, Levy D, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet*. 2009;373(9665):739–745.
5. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81(24):1879–1886.
6. Parikh NI, Pencina MJ, Wang TJ, et al. A risk score for predicting near-term incidence of hypertension: the Framingham heart study. *Ann Intern Med*. 2008;148(2):102–110.
7. Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *New Eng J Med*. 2008;359(21):2208–2219.
8. D'Agostino RB, Griffith JL, Schmidt CH, et al. Measures for evaluating model performance. *Proceedings of the Biometrics Section*. Alexandria, VA: American Statistical Association, Biometrics Section; 1997:253–258.
9. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *JAMA*. 2001;285(19):2486–2497.
10. Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New Eng J Med*. 2006;355(25):2631–2639.
11. Polonsky TS, McClelland RL, Jorgensen NW, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. *JAMA*. 2010;303(16):1610–1616.
12. Zethelius B, Berglund L, Sundström J, et al. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *New Eng J Med*. 2008;358(20):2107–2116.
13. Cook NR. Use and misuse of the receiver operating characteristics curve in risk prediction. *Circulation*. 2007;115(7):928–935.
14. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel cardiovascular risk: a scientific statement

from the American Heart Association. *Circulation*. 2009; 119(17):2408–2416.
15. Janes H, Pepe M, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med*. 2008;149(10):751–760.
16. Janssens ACJW, Little J, Ioannidis JPA, et al. Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. *Eur J Epidemiol*. 2011;26(4):255–259.
17. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–387.
18. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157–172.
19. Pencina MJ, D'Agostino RB Sr, Steyerberg E. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30(1):11–21.
20. Pencina MJ, D'Agostino RB Sr, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med*. 2012;31(2): 101–113.
21. Demler OV, Pencina MJ, D'Agostino RB. Equivalence of AUC improvement and significance of linear discriminant analysis coefficient under the assumptions of multivariate normality. *Stat Med*. 2011;30(12):1410–1418.
22. Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. *Stat Med*. 2010;29(24):2508–2520.
23. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
24. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J Educ Stat*. 1981;6(2):107–128.
25. Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159(9):882–890.
26. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Comments on integrated discrimination and net reclassification improvements: practical advice. *Stat Med*. 2008;27(2):207–212.
27. Mahalanobis PC. On the generalized distance in statistics. *Proc Natl Inst Sci India*. 1936;2:49–55.
28. Tzoulaki I, Liberopoulos G, Ioannidis JPA. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA*. 2009;302(21):2345–2352.
29. Pencina MJ, D'Agostino RB. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*. 2004; 23(13):2109–2123.
30. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950; 3(1):32–35.
31. Schnabel RB, Larson MG, Yamamoto JF, et al. Relations of biomarkers of distinct pathophysiological pathways and atrial fibrillation incidence in the community. *Circulation*. 2010; 121(2):200–207.
32. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574.
33. Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. *Med Decis Making*. 2008;28(1):146–149.