

COMMENTARY

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 (this issue)

M. S. Pepe^{*,†}, Z. Feng and J. W. Gu

Fred Hutchinson Cancer Research Center, Biostatistics and Biomathematics, 1100 Fairview Ave. N., M2-B500, Seattle, WA 98109, U.S.A.

1. INTRODUCTION

The evaluation of risk prediction markers has received much attention recently [1–3], and the paper by Pencina *et al.* [4] is an important contribution to this literature. There is a growing recognition that the receiver operating characteristic (ROC) curve, that has played a central role in evaluating diagnostic markers [5], has some serious limitations while evaluating risk prediction markers. In particular, a key attribute of the ROC curve is that it does not involve the original measurement scale for the marker. Rather, it provides a common scale on which the classification performances of different markers can be compared. This is incredibly valuable when the original measurement units are irrelevant as they typically are for diagnostic tests. However, for comparing risk models, with individuals’ risks represented as p_{new} versus p_{old} , the scales for p_{new} and p_{old} are the same and are very relevant to clinical application. Recent attempts to go beyond the ROC curve for evaluating risk prediction markers are concerned with taking the risk scale into account.

Pencina *et al.* [4] propose some new statistics to quantify the increment in performance when a new marker, Y , is added to an existing set of predictors X for predicting an outcome D . In their application, Y is HDL cholesterol and D is occurrence of a coronary heart disease (CHD) event within 10 years. It is assumed that both models, p_{new} and p_{old} , are well calibrated so that, to a good approximation, $p_{\text{new}} = P(D = 1|Y, X)$ and $p_{\text{old}} = P(D = 1|X)$. We use the subscript k below to denote old or new.

2. BEYOND SUMMARY INDICES TO PLOTS AND CURVES

Although their paper and most of my commentary discuss summary statistics, study results can rarely be reduced to a single number. What data display can be used to provide a fuller

*Correspondence to: M. S. Pepe, Fred Hutchinson Cancer Research Center, Biostatistics and Biomathematics, 1100 Fairview Ave. N., M2-B500, Seattle, WA 98109, U.S.A.

†E-mail: mspepe@u.washington.edu

description of the clinically pertinent information derived from the study? A scatterplot of p_{new} versus p_{old} seems like a natural complement to the summary statistics based on them, perhaps with different symbols or separate plots for cases (events) and controls (non-events). A line at $p_{\text{new}} = p_{\text{old}}$, in addition to horizontal and vertical lines at key risk thresholds, would allow one to see the extent and direction of change in risk induced by adding Y to the set of risk predictors. A summary index such as integrated discrimination improvement (IDI) or its components, integrated sensitivity (IS) and integrated 1-specificity (IP), cannot distinguish between a few large upward movements, a medium number of small upward movements, or a large number of movements in both upward and downward directions. Such important distinctions may be evident from a scatterplot. Granted, the two-way tabulations provided in Table II are essentially discretized versions of scatter plots. However, they rely heavily on choices of risk categories.

The marginal distributions of p_{new} and p_{old} have previously been proposed to compare the performances of two models (or markers). These can be plotted either for the population as a whole using predictiveness curves [2] or separately for cases and controls in an integrated plot [6]. The IDI summary statistic compares the means of these marginal distributions, $E(p_{\text{new}}|D=1)$ versus $E(p_{\text{old}}|D=1)$ in the IS component and $E(p_{\text{new}}|D=0)$ versus $E(p_{\text{old}}|D=0)$ in the IP component. Therefore, it can be regarded as a summary index for predictiveness and integrated plots. Plotting the entire distribution provides additional information. One can see, for example, the proportions of cases (or controls) whose risks lie above (or below) various thresholds for models with and without the new marker. This is similar to comparing the margins of Table II, but using a continuous rather than a discretized scale.

The two-way scatterplot or tabulation also allows one to view the distribution of p_{new} stratified on p_{old} , i.e. for the population whose risks calculated from baseline covariates, X , are within an interval. Cook [3] summarizes incremental performance by tabulating the conditional distribution of p_{new} conditional on p_{old} being in the medium-risk category. By considering p_{old} as a covariate, the methods of Huang *et al.* [2] for estimating covariate-specific predictiveness curves could be employed to arrive at continuous versions of these conditional distributions.

3. MANY INTERPRETATIONS FOR THE IDI STATISTIC

The IDI index proposed by Pencina *et al.* is motivated by considering improvement in mean risk for case and control populations:

$$\text{IDI} = E\{(p_{\text{new}} - p_{\text{old}})|D=1\} - E\{(p_{\text{new}} - p_{\text{old}})|D=0\} \quad (1)$$

Equivalently, the authors express it as the change in the mean risk difference, $\text{MRD}_k = E\{p_k|D=1\} - E\{p_k|D=0\}$, also called the discrimination slope:

$$\text{IDI} = \text{MRD}_{\text{new}} - \text{MRD}_{\text{old}} \quad (2)$$

Since the mean of a non-negative random variable, W , can be expressed in terms of its survivor function [7, Section 21], $E(W) = \int P(W > w) dw$, expressions in terms of integrated sensitivity and specificity follow from (1):

$$\text{IDI} = [\text{IS}_{\text{new}} - \text{IS}_{\text{old}}] + [\text{IP}_{\text{old}} - \text{IP}_{\text{new}}] \quad (3)$$

where $\text{IS}_k = \int P(p_k > c|D=1) dc$ is integrated true-positive fraction (TPF = sensitivity) and $\text{IP}_k = \int P(p_k > c|D=0) dc$ is integrated false-positive fraction (FPF = 1 - specificity). The authors note

that this is the change in Youden's index, TPF – FPF, integrated uniformly over (0, 1):

$$\text{IDI} = \int \text{YI}_{\text{new}}(c) \, dc - \int \text{YI}_{\text{old}}(c) \, dc \quad (4)$$

where $\text{YI}_k(c) = P(p_k > c | D = 1) - P(p_k > c | D = 0)$ is Youden's index for the binary rule, $p_k > c$.

We now provide some additional interpretations for IDI that are not in Pencina *et al.*'s paper. Let PEV be the proportion of the explained variation, a generalization of R^2 from linear to binary regression [8]:

$$\begin{aligned} \text{PEV}_{\text{old}} &\equiv \frac{\text{var}(D) - E\{\text{var}(D|X)\}}{\text{var}(D)} \\ &= \text{var}(p_{\text{old}})/\rho(1 - \rho) \end{aligned}$$

where $\rho = P(D = 1)$ and the equality follows because $E(D|X) = P(D = 1|X) = p_{\text{old}}$. Similarly, we write

$$\begin{aligned} \text{PEV}_{\text{new}} &\equiv \frac{\text{var}(D) - E\{\text{var}(D|X, Y)\}}{\text{var}(D)} \\ &= \text{var}(p_{\text{new}})/\rho(1 - \rho) \end{aligned}$$

Hu *et al.* [8] note another representation as $\text{PEV}_k = \text{corr}(D, p_k)$.

The following, rather unintuitive, result is proven in the Appendix:

$$\text{IDI} = \text{PEV}_{\text{new}} - \text{PEV}_{\text{old}} \quad (5)$$

The above states that IDI is the change in R^2 achieved by adding the new marker to the binary regression risk model. This puts IDI in a very traditional and familiar framework.

Yet another interpretation for IDI is in terms of classification error or absolute residuals. The absolute residuals $|D - p_k|$ are $(1 - p_k)$ for cases and p_k for controls. Let us take a weighted average with weights for cases, $w_1 = \rho^{-1}$ and weights for controls, $w_0 = (1 - \rho)^{-1}$. Then

$$\begin{aligned} E\{\text{residual}\}_k &= w_1 E(1 - p_k | D = 1)\rho + w_0 E(p_k | D = 0)(1 - \rho) \\ &= E(1 - p_k | D = 1) + E(p_k | D = 0) \end{aligned}$$

It follows that

$$\text{IDI} = E\{\text{residual}\}_{\text{old}} - E\{\text{residual}\}_{\text{new}} \quad (6)$$

We conclude that the IDI statistic is a general and natural measure of the improvement in prediction afforded by adding Y to the risk model. It can be motivated in at least six different ways including from the point of view of traditional regression methods: as changes in mean absolute residuals, equation (6), and as changes in R^2 , equation (5). The interpretations provided by Pencina *et al.*, as average improvements in risk values for cases and controls ((1) and (2)) as well as changes in integrated sensitivity and specificity ((3) and (4)) are compelling and easy to understand. Having multiple interpretations for IDI is not just academically interesting. It implies that approaches to analyses that are apparently different are in fact the same. This is reassuring.

4. INFERENCE FOR IDI AND ITS COMPONENTS

In classic ROC analysis, the sensitivity of a marker is a property entirely of the cases and specificity is a property of controls. However, with risk prediction markers, cases and controls both enter into the risk model. This induces some interesting relationships between sensitivity and specificity calculated on the basis of risk. In this article, the two components of IDI are

$$IS_{\text{new}} - IS_{\text{old}} = E[p_{\text{new}}|D=1] - E[p_{\text{old}}|D=1]$$

and

$$IP_{\text{old}} - IP_{\text{new}} = E[p_{\text{old}}|D=0] - E[p_{\text{new}}|D=0]$$

Since

$$\begin{aligned} \rho = P[D=1] &= E[p_k] = \rho E[p_k|D=1] + (1-\rho)E[p_k|D=0] \\ &= \rho IS_k + (1-\rho)IP_k \end{aligned}$$

subtracting equations for $E[p_{\text{new}}]$ and $E[p_{\text{old}}]$ implies that

$$0 = \rho(IS_{\text{new}} - IS_{\text{old}}) + (1-\rho)(IP_{\text{new}} - IP_{\text{old}})$$

That is, there is a fundamental relationship between the two components of IDI

$$IS_{\text{new}} - IS_{\text{old}} = \frac{(1-\rho)}{\rho}(IP_{\text{old}} - IP_{\text{new}}) \quad (7)$$

at least in large samples and assuming well-calibrated models. There are two important implications. First, an improvement in sensitivity must be accompanied by an improvement in specificity. One cannot technically fix the specificity improvement at 0 and investigate improvements in sensitivity, as implied in the paper. Indeed, a test for improvement in IS is equivalent to a test for improvement in IP, since one is a scaled version of the other. Second, in settings where $\rho = P[D=1]$ is small, equation (7) implies that the improvement in sensitivity will be much larger than the improvement in specificity. For example, in the CHD application, $\rho = \frac{183}{3264} = 5.6$ per cent, so that the improvement in specificity is 0.059 times the improvement in sensitivity. This relative improvement is borne out in the example.

I am not convinced that statistical tests of the null hypothesis, $H_0 : \text{IDI} = 0$, are particularly useful. In fact, one can show that testing the null hypothesis, $H_0 : \text{IDI} = 0$, is equivalent to testing the null hypothesis, $H_0 : \beta = 0$, where β is the regression coefficient for Y in a risk model that includes (X, Y) as predictors (see Appendix). This follows intuitively from the representation of IDI in terms of the change in R^2 , equation (5). Pencina *et al.* argue that tests of $H_0 : \beta = 0$ are often significant even when improvements in model performance are minimal. In a similar vein, we conjecture that tests of the null hypothesis, $H_0 : \text{IDI} = 0$, are likely to be statistically significant even when IDI is extremely small. Such is the case in their application. Instead of testing hypotheses, perhaps the focus should be on providing a confidence interval for the summary measure of improvement. The variance expression provided as the denominator of equation (15) in Pencina *et al.*'s paper, however, may be an under estimate. It does not appear to account for variability in \hat{p}_k due to sampling variability in the regression coefficient estimates. This additional source of variability could be incorporated with bootstrap resampling.

5. INSIGHTS REGARDING AUC AND RELATIONSHIPS WITH (TPF, FPF)

Although the area under the ROC curve is often reported, it has been widely criticized [3, 5]. Pencina *et al.* provide its common interpretation as ‘the probability that given two subjects, one who will develop an event and one who will not, the model will assign a higher probability of an event to the former.’ However, subjects are never presented in pairs in clinical practice. This interpretation does not appear to be clinically relevant.

Their insightful discussion of the relationship between IS and the AUC suggests a broader class of performance measures within which these and other performance measures fall. In practice, different risk thresholds C^H may be used to designate a subject as ‘high risk’, i.e. to choose the corresponding intervention. Individual patients and their caregivers have different tolerances for risk. Variability in personal and financial resources affects risk tolerance. Moreover, perceived costs and benefits can affect a person’s choice of risk threshold. Let $F^H(c)$ be the cumulative probability distribution of high-risk thresholds, C^H , used in practice. Define TPF^H as the probability that a subject destined to have an event has a high-risk designation:

$$\text{TPF}_k^H = P(p_k > C^H | D = 1) = \int_0^1 P(p_k > c | D = 1) dF^H(c) \quad (8)$$

and similarly define FPF^H as the probability that a control, destined not to have an event, is classified as high risk:

$$\text{FPF}_k^H = P(p_k > C^H | D = 0) = \int_0^1 P(p_k > c | D = 0) dF^H(c) \quad (9)$$

As noted by Pencina *et al.*, if the probability distribution of C^H is uniform on $(0, 1)$, then $\text{TPF}_k = \text{IS}_k$ and $\text{FPF}_k = \text{IP}_k$. However, in practice, the thresholds used for high-risk designation are unlikely to be uniformly distributed over the entire $(0, 1)$ domain. In fact, in the cardiovascular disease applications presented here and elsewhere, it appears that F^H might have a point mass at $C^H = 0.20$. If that were really so, then the values $\text{TPF}_{\text{new}}^H = 0.191$ versus $\text{TPF}_{\text{old}}^H = 0.131$ and $\text{FPF}_{\text{old}}^H = 0.032$ versus $\text{FPF}_{\text{new}}^H = 0.033$ reported in their application would be sufficient.

Pencina *et al.* also note that with the high-risk threshold distribution chosen to equal the risk distribution in controls, i.e. $F^H(c) = P(p_k < c | D = 0) = \text{specificity}(c)$, we have $\text{TPF}_k^H = \text{AUC}_k$ and $\text{FPF}_k^H = 0.5$. That is, if the distribution of high-risk thresholds used in practice were such that associated false-positive rates were uniformly distributed in $(0, 1)$, then the proportion of cases classified as high risk, i.e. the marginal or net sensitivity, would be given by AUC and half of the controls would be classified as high risk. Again this is not a realistic scenario, leaving in doubt the value of the AUC for practical evaluation of a marker.

In cancer-screening research, we often choose a single-marker threshold corresponding to a very low FPF, f_0 , since maintaining a very low FPF is necessary in order to avoid large numbers of healthy people undergoing unnecessary work-up procedures [9]. That scenario corresponds to a point mass for $F^H(c)$ at the $1 - f_0$ quantile of the control distribution and TPF^H and FPF^H are recognized as $\text{ROC}(f_0)$ and f_0 , respectively. Choosing a high-risk threshold distribution to yield a uniform distribution of false-positive rates over $(0, f_0)$ yields TPF^H equal to the partial AUC [10].

Low-risk designation counterparts of TPF and FPF can be defined. Let C^L be an individual’s low-risk threshold. The proportion of cases that receive a low-risk designation, with model k ,

$k = \text{new, old, is}$

$$1 - \text{TPF}_k^L = P(p_k < C^L | D = 1) = \int_0^1 P(p_k < c | D = 1) dF^L(c) \quad (10)$$

and the corresponding proportion of controls is

$$1 - \text{FPF}_k^L = P(p_k < C^L | D = 0) = \int_0^1 P(p_k < c | D = 0) dF^L(c) \quad (11)$$

where F^L is an assumed population distribution of thresholds for designating subjects as low risk. Again it seems sensible to consider the population distribution of low-risk thresholds likely to be used in practice in evaluating risk prediction markers.

Comparisons of $\text{TPF}_{\text{new}}^H$ with $\text{TPF}_{\text{old}}^H$, $1 - \text{TPF}_{\text{new}}^L$ with $1 - \text{TPF}_{\text{old}}^L$, $\text{FPF}_{\text{new}}^H$ with $\text{FPF}_{\text{old}}^H$, and $1 - \text{FPF}_{\text{new}}^L$ with $1 - \text{FPF}_{\text{old}}^L$ provide a simple and meaningful basis for evaluation. Indeed, putting these together with incidence, ρ , and costs (or more generally losses or utilities) associated with incorrect risk designations would allow a comprehensive approach to evaluate risk prediction markers [1].

6. RISK CATEGORIES AND RECLASSIFICATION

The net reclassification improvement (NRI) summary index is introduced for settings where risk categories are defined. In cardiovascular disease research, there is some consensus on what constitutes meaningful risk categories. The current paper uses the categories 0–6, 6–20 and 20–100 per cent 10-year risk of a CHD events. Presumably, the category thresholds are based on weighing the net cost *versus* net benefit of incorrect *versus* correct designation. In particular, suppose that treating a subject as high-risk has a net benefit B^+ if he is destined to have an event in the absence of the treatment and a net cost B^- if he is not destined to have an event. A subject whose risk of an event is p has an expected benefit

$$E\{B(p)\} = pB^+ - (1 - p)B^- \quad (12)$$

A natural choice for the high-risk threshold is $p : p/(1 - p) = B^-/B^+$, where the expected benefit crosses 0. It would be interesting to know whether and how such considerations were formally involved in choosing the 20 and 6 per cent risk thresholds in CHD research. Researchers in other medical fields such as cancer need to develop consensus about meaningful risk categories and could perhaps learn from our colleagues in cardiovascular disease research.

Note that an individual considering having a risk prediction marker measured will ultimately be required to make two decisions. After his marker value is available, his risk can be calculated and he will formally or informally assess the expected benefit of treatment for him in his decision to the avail of treatment. However, the preliminary decision is whether or not to have his marker measured in the first place. This decision is also based on an assessment of the expected benefit, given information available to him, which at this point is not based on his risk but upon knowledge about disease incidence and the probabilities of his receiving a high-risk designation if he is destined to have an event or not. His expected benefit of *testing* is

$$E\{B(\text{test})\} = \rho(\text{TPF})B^+ - (1 - \rho)(\text{FPF})B^- \quad (13)$$

COMMENTARY

Table II'. Overall risk reclassification.

Risk without HDL	Risk with HDL			Total	Per cent reclassified
	<6 per cent	6–20 per cent	>20 per cent		
<6 per cent	1998	157	0	2155	7.2
6–20 per cent	152	790	45	987	20.0
>20 per cent	1	28	93	122	23.8
Total	2151	975	138		

The point is that, although the high-risk threshold might be chosen based on (12), evaluation of the testing strategy depends on $(\text{TPF}^H, \text{FPF}^H)$, as defined in (8) and (9). The risk threshold distribution F^H may be an individual-specific point mass when an individual makes the decision to be tested or not. On the other hand, the risk distribution would pertain to practice in the population when evaluating the impact of testing on the population.

Consider again Table II of Pencina *et al.*'s paper that displays a cross-tabulation of risk categories with and without the HDL marker. Cook [3] considers such a tabulation but ignores the true outcome shown as Table II' here. Cook quantifies the impact of the marker with the proportion of subjects in the middle-risk category without HDL who are reclassified into low- or high-risk categories by the model that includes the new marker. From Table II' this is 20.0 per cent, a seemingly large effect. However, this analysis ignores the first and third rows of Table II' that show substantial numbers of subjects reclassified from high- and low-risk categories *into* the middle-risk category. The *net* movement out of the middle-risk category, seen from the margins of the table, is only $987 - 975 = 12$ subjects. In contrast to Cook [3], Pencina *et al.* took the approach of evaluating *net* improvement with the NRI index.

They also argue convincingly that risk reclassifications must be gauged relative to the true outcome. Therefore, their tabulation in Table II separates subjects accordingly. We have argued similarly [11]. Risk reclassifications are improvements only if they are in the upward direction for cases and in the downward direction for controls.

However, I suspect that not all upward reclassifications are equivalent for cases, and that not all downward reclassifications are equivalent for controls. For example, for cases, movement from the low-risk category to the medium-risk category may not be so important as movement from the medium- to the high-risk category. A concern we have about the NRI measure is that it treats all reclassifications that are in the correct direction as equivalent. An alternative is to report the components of NRI, namely changes in $\text{TPF}^H = P[\text{high-risk designation}|D = 1]$, $\text{FPF}^H = P[\text{high-risk designation}|D = 0]$, $1 - \text{TPF}^L = P[\text{low-risk designation}|D = 1]$ and $1 - \text{FPF}^L = P[\text{low-risk designation}|D = 0]$. The components are reported in their example in Table II.

7. CONCLUSIONS

Risk modeling is a ubiquitous exercise in clinical and epidemiological research. Risk models are sufficient for a subject who wishes to calculate his risk given his predictors. Relative risk associated with risk factors are key in etiological research. However, to evaluate the population impact of a marker as a risk predictor, the population distribution of the marker and other predictors must be considered in addition to the risk model itself. The population distribution of risks depends on

COMMENTARY

both the risk model and on the predictor distributions. Pencina *et al.* are to be congratulated for suggesting a sensible approach to evaluating the impact of a new marker by comparing population risk distributions with and without the marker, separately for cases and controls. They propose summary statistics, although modification might be considered in the context of specific applications to incorporate probability densities for risk thresholds that are likely to be used in practice. It is also important to develop data displays that provide more complete information than can be contained in summary statistics.

We are still at the point of developing conceptual approaches for proper analysis of risk prediction markers. Most of our comments here ignore sampling variability and assume that estimated risks are good approximations to true risks. Development of techniques for estimation and inference will keep methodological statisticians busy in the future, but first we must come to a consensus about the appropriate conceptual approaches. The Pencina *et al.* paper is an important step towards achieving that goal.

APPENDIX

Proof of (5)

Let W denote the predictors in the model. With subscript $k = \text{old}$ the predictors are X and when $k = \text{new}$ we have $W = (X, Y)$. We write $p(W)$ for the risk based on W . Consider

$$\begin{aligned}
 E(p(W)|D=1) &= \int p(W) dF(W|D=1) \\
 &= \int \frac{p(W)P(D=1|W)}{P(D=1)} dF(W) \\
 &= \rho^{-1} \int p^2(W) dF(W) \\
 &= \rho + (1 - \rho) \left\{ \frac{\int p^2(W) dF(W) - \rho^2}{\rho(1 - \rho)} \right\} \\
 &= \rho + (1 - \rho) \text{var}(p(W))/\rho(1 - \rho) \\
 &= \rho + (1 - \rho)\text{PEV}
 \end{aligned}$$

Since $\rho = P(D=1) = E(p(W)) = \rho E(p(W)|D=1) + (1 - \rho)E(p(W)|D=0)$, we have

$$\begin{aligned}
 E(p(W)|D=0) &= \rho\{1 - E(p(W)|D=1)\}/(1 - \rho) \\
 &= \rho\{1 - \text{PEV}\}
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \text{MRD} &\equiv E(p(W)|D=1) - E(p(W)|D=0) \\
 &= \rho + (1 - \rho)\text{PEV} - \rho + \rho\text{PEV} \\
 &= \text{PEV}
 \end{aligned}$$

The result then follows from equation (2).

COMMENTARY

Equivalence of two null hypotheses

Write $H_0 : \text{IDI} = 0$ and $H_0^* : p_{\text{new}} = P(D = 1|X, Y) = P(D = 1|X) = p_{\text{old}}$. Clearly, if H_0^* holds then H_0 holds. If H_0 holds then $\text{var}(p_{\text{new}}) = \text{var}(p_{\text{old}})$ according to equation (5). But

$$\begin{aligned}\text{var}(p_{\text{new}}) &= \text{var}(P(D = 1|X, Y)) \\ &= E \text{var}\{P(D = 1|X, Y)|X\} + \text{var} E\{P(D = 1|X, Y)|X\} \\ &= E \text{var}\{P(D = 1|X, Y)|X\} + \text{var}\{P(D = 1|X)\} \\ &= E \text{var}\{P(D = 1|X, Y)|X\} + \text{var}(p_{\text{old}})\end{aligned}$$

Therefore, under H_0 , $E \text{var}\{P(D = 1|X, Y)|X\} = 0$, which implies that

$$P(D = 1|X, Y) = E\{P(D = 1|X, Y)|X\} = P(D = 1|X)$$

with probability 1.

ACKNOWLEDGEMENT

Support for this research was provided by the National Institutes of Health under Contract numbers RO1 GM054438 and UO1 CA086368.

REFERENCES

1. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005; **6**:227–239.
2. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics* 2007, in press.
3. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115**:928–935.
4. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, in this issue. DOI: 10.1002/sim.2929.
5. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute* 2003; **95**:511–515.
6. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *UW Biostatistics Working Paper Series*, 2006; Working Paper 289. Available at: <http://www.bepress.com/uwbiostat/paper289>.
7. Billingsley P. *Probability and Measure*. Wiley: New York, 1979.
8. Hu B, Palta M, Shao S. Properties of R^2 statistics for logistic regression. *Statistics in Medicine* 2006; **25**: 1383–1395.
9. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson M, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of disease. *Journal of the National Cancer Institute* 2001; **93**:1054–1061.
10. McClish DK. Analysing a portion of the ROC curve. *Medical Decision Making* 1989; **9**:190–195.
11. Pepe MS, Janes H, Gu W. Letter to the editor in response to: Cook NR 'Use and misuse of the receiver operating characteristic curve in risk prediction'. *Circulation*, to appear.