STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:157–172
Published online 13 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2929

Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond

Michael J. Pencina[1,*,†], Ralph B. D'Agostino Sr[1], Ralph B. D'Agostino Jr[2] and Ramachandran S. Vasan[3]

[1]*Department of Mathematics and Statistics, Framingham Heart Study, Boston University, 111 Cummington St., Boston, MA 02215, U.S.A.*
[2]*Department of Biostatistical Sciences, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, U.S.A.*
[3]*Framingham Heart Study, Boston University School of Medicine, 73 Mount Wayte Avenue, Suite 2, Framingham, MA 01702-5803, U.S.A.*

SUMMARY

Identification of key factors associated with the risk of developing cardiovascular disease and quantification of this risk using multivariable prediction algorithms are among the major advances made in preventive cardiology and cardiovascular epidemiology in the 20th century. The ongoing discovery of new risk markers by scientists presents opportunities and challenges for statisticians and clinicians to evaluate these biomarkers and to develop new risk formulations that incorporate them. One of the key questions is how best to assess and quantify the improvement in risk prediction offered by these new models. Demonstration of a statistically significant association of a new biomarker with cardiovascular risk is not enough. Some researchers have advanced that the improvement in the area under the receiver-operating-characteristic curve (AUC) should be the main criterion, whereas others argue that better measures of performance of prediction models are needed. In this paper, we address this question by introducing two new measures, one based on integrated sensitivity and specificity and the other on reclassification tables. These new measures offer incremental information over the AUC. We discuss the properties of these new measures and contrast them with the AUC. We also develop simple asymptotic tests of significance. We illustrate the use of these measures with an example from the Framingham Heart Study. We propose that scientists consider these types of measures in addition to the AUC when assessing the performance of newer biomarkers. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: discrimination; model performance; AUC; risk prediction; biomarker

[*] Correspondence to: Michael J. Pencina, Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, U.S.A.

[†] E-mail: mpencina@bu.edu

Contract/grant sponsor: National Heart, Lung, and Blood Institute's Framingham Heart Study; contract/grant numbers: N01-HC-25195, 2K24 HL 04334

Received 4 April 2007
Accepted 5 April 2007

INTRODUCTION

Over 30 years after the construction of the first multivariable risk prediction model (also called risk profile model) predicting the probability of developing cardiovascular disease (CVD) [1], researchers continue to seek new risk factors that can predict CVD and that can be incorporated into risk assessment algorithms. A general consensus exists that the information regarding an individual's age, baseline levels of systolic and diastolic blood pressure and serum cholesterol, smoking and diabetes status are all useful predictors of the CVD risk over a reasonable time period in the future, typically over 1–10 years [2–5]. Quantification of vascular risk is accomplished through risk equations or risk score sheets that have been developed on the basis of observations from large cohort studies [2–6]. For example, the Framingham risk score has been routinely applied, validated and calibrated for use in different countries, and for different ethnicities across countries [7–9]. Various statistical models have been utilized over the decades to develop these equations. Presently, Cox proportional hazards and Weibull parametric models seem to be among the most frequently used ones [2–6]. However, CVD risk prediction is an ongoing work in progress. New risk factors or markers are being identified and proposed constantly, and vie with each other for consideration for incorporation into risk prediction algorithms. Plasminogen-activator inhibitor type 1, gamma glutamyl transferase, C-reactive protein (CRP), B-type natriuretic peptide, urinary albumin-to-creatinine ratio, left ventricular hypertrophy or fibrinogen are only a few examples from a very long list [2, 10–12].

The critical question arises as to how to evaluate the usefulness of a new marker. D'Agostino [13] lists four initial decisions that guide the process: (1) defining the population of interest; (2) defining the outcome of interest; (3) choosing how to incorporate the competing pre-existing set of risk factors and (4) selecting the appropriate model and tests to evaluate the incremental yield of a new biomarker. This paper focuses on the last issue, assuming that we have adequately defined or answered the issues described in (1)–(3).

The most basic necessary condition required of any new marker is its statistical significance. It is hard to imagine that one would argue for an inclusion of a new marker into a risk prediction formulation if it is not related to the outcome of interest in a statistically significant manner. Statistical significance, however, does not imply either clinical significance or improvement in model performance. Indeed, many biomarkers with weak or moderate relations to the outcome of interest can be associated in a statistically significant fashion if examined using a large enough sample size.

Evaluation of risk prediction models and adjustments to them require model performance measures [7, 14]. A key measure of the clinical utility of a survival model is its ability to discriminate or separate those who will develop the event of interest from those who will not. Various measures have been proposed to capture discrimination [14], but the area under the receiver-operating-characteristic (ROC) curve (AUC) is the most popular metric [14–16]. Its probabilistic interpretation addresses simply and directly its discriminatory ability. It is the probability that given two subjects, one who will develop an event and the other who will not, the model will assign a higher probability of an event to the former. Its traditional application in the context of binary outcomes has been extended to the time-to-event models, which are the standard models for CVD risk prediction [17, 18].

Researchers, extending existing methodology, began evaluating new markers based on their ability to increase the AUC. It quickly became apparent that, for models containing standard risk factors and possessing reasonably good discrimination, very large 'independent' associations of

DOI: 10.1002/sim

the new marker with the outcome are required to result in a meaningfully larger AUC [19–21]. None of the numerous new markers proposed comes close in magnitude to these necessary levels of association. In response to this, some scientists have argued that we need to wait for new and better markers; other researchers have sought model performance measures beyond the AUC to evaluate the usefulness of markers. Reassignment of subjects into risk categories (reclassification tables) and predictiveness curves form opposite ends of the spectrum of new ideas [22–24]. These efforts address Greenland's and O'Malley's suggestion that 'statisticians should seek new ways, beyond the ROC curve, to evaluate clinically useful new markers for their ability to improve upon current models such as the Framingham Risk Score' [20].

In this paper, we propose two new ways of evaluating the usefulness of a new marker. They fall somewhere in the middle of the spectrum mentioned above. One is based on event-specific reclassification tables, and the other on the new model's ability to improve integrated (average) sensitivity (IS) without sacrificing integrated (average) specificity.

We start with a careful look at reclassification tables and suggest an objective way of quantifying improvement in categories through what we term 'the net reclassification improvement' or NRI. This method requires that there exist *a priori* meaningful risk categories (e.g. 0–6, 6–20, >20 per cent 10-year risk of coronary heart disease [CHD] based on the Third Adult Treatment Panel [ATP III] risk classification [5]). Then, we extend this idea to the case of no 'cut-offs' to define our second measure, the integrated discrimination improvement (IDI). We propose sample estimators for both measures and show how the IDI can be estimated by the difference in discrimination slopes proposed by Yates [25]. We also derive simple asymptotic tests to determine whether the improvements in our measures are significantly different from zero. As an illustration, we show a Framingham Heart Study example, in which the NRI and IDI indicate that HDL cholesterol offers statistically significant improvement in the performance of a CHD model even though no meaningful or significant improvement in the AUC is observed. Formal mathematical developments for some identities are presented in the Appendix.

NET RECLASSIFICATION IMPROVEMENT AND INTEGRATED DISCRIMINATION IMPROVEMENT

In this section, we propose two new ways of assessing improvement in model performance offered by a new marker. The NRI focuses on reclassification tables constructed separately for participants with and without events, and quantifies the correct movement in categories—upwards for events and downwards for non-events. The IDI does not require categories, and focuses on differences between ISs and 'one minus specificities' for models with and without the new marker. This section introduces the concepts in general terms; a more formal discussion is presented in the next section followed by a practical example.

From AUC to reclassification

The developments concerning the AUC come from applications to diagnostic testing in radiology [16]. AUC can be defined as the area under the plot of sensitivity *vs* 'one minus specificity' for all possible cut-off values. This definition has been shown to be equivalent to defining AUC as the probability that a given diagnostic test (or predictive model in our case) assigns a higher probability of an event to those who actually have (or develop) events [16]. The improvement in AUC for

DOI: 10.1002/sim

a model containing a new marker is defined simply as the difference in AUCs calculated using a model with and without the marker of interest. This increase, however, is often very small in magnitude; for example, Wang *et al*. show that the addition of a biomarker score to a set of standard risk factors predicting CVD increases the model AUC only from 0.76 to 0.77 [12]. Ware and Pepe show simple examples in which enormous odds ratios are required to meaningfully increase the AUC [19, 21].

Because of the above, some researchers started looking at different methods of quantifying the improvement. Reclassification tables have been gaining popularity in medical literature [10, 22, 23]. For example, Ridker *et al*. [10] compare a model developed for CVD risk prediction in women using only standard risk factors ('old' model) with a model that also includes parental history of myocardial infarction and CRP ('new' model), and observe a minimal increase in the AUC from 0.805 to 0.808. However, when they classify the predicted risks obtained using their two models (old and new) into four categories (0–5, 5–10, 10–20, >20 per cent 10-year CVD risk) and then cross-tabulate these two classifications, they show that about 30 per cent of individuals change their category when comparing the new model with the old one. They further calculate the actual event rates for those reclassified and call the reclassification successful if the actual rate corresponds to the new model's category.

Unfortunately, reclassification tables constructed and interpreted in this manner offer limited means of evaluating improvement in performance. Relying solely on the number or percentage of subjects who are reclassified can be misleading. Additionally, calculating event rates among the reclassified individuals does not lead to an objective assessment of the true improvement in classification. For instance, even if we reclassify 100 people from the 10–20 per cent 10-year CVD risk category into the above 20 per cent group and the 'actual' event rate among these individuals is 25 per cent, we improved the placement of 25 people, but not the remaining 75 who should have stayed in the lower risk category.

We suggest a different way of constructing and interpreting the reclassification tables. The reclassification of people who develop and who do not develop events should be considered separately. Any 'upward' movement in categories for event subjects (i.e. those with the event) implies improved classification, and any 'downward movement' indicates worse reclassification. The interpretation is opposite for people who do not develop events. The improvement in reclassification can be quantified as a sum of differences in proportions of individuals moving up minus the proportion moving down for people who develop events, and the proportion of individuals moving down minus the proportion moving up for people who do not develop events. We call this sum the NRI. Equivalently, the NRI can be calculated by computing the difference between the proportions of individuals moving up and the proportion of individuals moving down for those who develop events and the corresponding difference in proportions for those who not develop events, and taking a difference of these two differences. A simple asymptotic test that can be used to determine the significance of the improvement, separately for event and non-event individuals and combining the two groups (NRI), is presented in the next section.

From reclassification to discrimination slopes

One potential drawback of the reclassification-based measure defined above is its dependence on the choice of categories. This limitation can be overcome by further extending the concept of the NRI. If we assign 1 for each upward movement, -1 for each downward movement and 0 for no

DOI: 10.1002/sim

movement in categories, the NRI can be expressed as

$$\frac{\sum_{i \text{ in events}} v(i)}{\# \text{ events}} - \frac{\sum_{j \text{ in nonevents}} v(j)}{\# \text{ nonevents}} \tag{1}$$

where $v(i)$ is the above-defined movement indicator. Now consider the categorization so fine that each person belongs to their own category. Then any increase in predicted probabilities for individuals with events means upward movement ($v(i)=1$) and any decrease is a downward movement ($v(i)=-1$). In this case, it makes sense to assign to each person the actual difference in predicted probabilities instead of 1, -1 or 0. If we denote the new model-based predicted probabilities of an event by $\hat{p}_{\text{new}}$ and old model-based probabilities by $\hat{p}_{\text{old}}$, we have

$$\frac{\sum_{i \text{ in events}} (\hat{p}_{\text{new}}(i) - \hat{p}_{\text{old}}(i))}{\# \text{ events}} - \frac{\sum_{j \text{ in nonevents}} (\hat{p}_{\text{new}}(j) - \hat{p}_{\text{old}}(j))}{\# \text{ nonevents}} \tag{2}$$

We show in the Appendix that the first term in (2) quantifies improvement in sensitivity and the negative of the second term quantifies improvement in specificity. Also, by rearranging the terms in (2), we observe that it is equivalent to the difference in discrimination slopes as introduced by Yates [25, 26] (discrimination slope can be defined as a difference between mean predicted probabilities of an event for those with events and the corresponding mean for those without events).

The difference in model-based discrimination slopes is an important measure of improvement in model performance. As shown in the Appendix, it is a sample equivalent of the difference between the integrated difference in sensitivities and the integrated difference in 'one minus specificities' between the new and old models. This integration is over all possible cut-offs. Thus, it quantifies jointly the overall improvement in sensitivity and specificity. In simpler terms, the area under the sensitivity curve is estimated by the mean of predicted probabilities of an event for those who experience events, and the area under the 'one minus specificity' curve is estimated by the mean of predicted probabilities of an event for those who do not experience events. We suggest the integrated differences in sensitivities and 'one minus specificities' and their difference as another measure of improvement in performance offered by the new marker. We call the last difference the IDI and estimate it using the difference in discrimination slopes. A simple asymptotic test of significance is provided in the next section.

STATISTICAL PROCEDURES AND CONSIDERATIONS

Net reclassification improvement

Consider a situation in which predicted probabilities of a given event of interest are estimated using two models that share all risk factors, except for one new marker. Let us categorize the predicted probabilities based on these two models into a set of clinically meaningful ordinal categories of absolute risk and then cross-tabulate these two classifications. Define upward movement (up) as a change into higher category based on the new model and downward movement (down) as a change in the opposite direction. If D denotes the event indicator, we define the NRI as

$$\text{NRI} = [P(\text{up}|D=1) - P(\text{down}|D=1)] - [P(\text{up}|D=0) - P(\text{down}|D=0)] \tag{3}$$

To estimate NRI using sample data, we define estimators for the four probabilities comprising the NRI:

$$\hat{P}(\text{up}|D=1)=\hat{p}_{\text{up,events}}=\frac{\text{\# events moving up}}{\text{\# events}} \tag{4}$$

$$\hat{P}(\text{down}|D=1)=\hat{p}_{\text{down,events}}=\frac{\text{\# events moving down}}{\text{\# events}} \tag{5}$$

$$\hat{P}(\text{up}|D=0)=\hat{p}_{\text{up,nonevents}}=\frac{\text{\# nonevents moving up}}{\text{\# nonevents}} \tag{6}$$

$$\hat{P}(\text{down}|D=0)=\hat{p}_{\text{down,nonevents}}=\frac{\text{\#nonevents moving down}}{\text{\# nonevents}} \tag{7}$$

The NRI is estimated as

$$\widehat{\text{NRI}}=(\hat{p}_{\text{up,events}}-\hat{p}_{\text{down,events}})-(\hat{p}_{\text{up,nonevents}}-\hat{p}_{\text{down,nonevents}}) \tag{8}$$

We note that formula (8) is equivalent to formula (1) of the previous section.

Assuming independence between event and non-event individuals and following McNemar's [27] logic for significance testing in correlated proportions (and using the properties of multinomial distribution), we arrive at a simple asymptotic test for the null hypothesis of NRI $=0$:

$$z=\frac{\widehat{\text{NRI}}}{\sqrt{\dfrac{\hat{p}_{\text{up,events}}+\hat{p}_{\text{down,events}}}{\text{\# events}}+\dfrac{\hat{p}_{\text{up,nonevents}}+\hat{p}_{\text{down,nonevents}}}{\text{\# nonevents}}}} \tag{9}$$

Individual components of the NRI, assessing improvement in event and non-event classifications, can be tested using

$$z_{\text{events}}=\frac{\hat{p}_{\text{up,events}}-\hat{p}_{\text{down,events}}}{\sqrt{\dfrac{\hat{p}_{\text{up,events}}+\hat{p}_{\text{down,events}}}{\text{\# events}}}} \tag{10}$$

$$z_{\text{nonevents}}=\frac{\hat{p}_{\text{down,nonevents}}-\hat{p}_{\text{up,nonevents}}}{\sqrt{\dfrac{\hat{p}_{\text{down,nonevents}}+\hat{p}_{\text{up,nonevents}}}{\text{\# nonevents}}}} \tag{11}$$

Integrated discrimination improvement

Denote by IS the integral of sensitivity over all possible cut-off values from the (0, 1) interval and by IP the corresponding integral of 'one minus specificity'. We define the IDI as follows:

$$\text{IDI}=(\text{IS}_{\text{new}}-\text{IS}_{\text{old}})-(\text{IP}_{\text{new}}-\text{IP}_{\text{old}}) \tag{12}$$

Subscript 'new' in the above expression refers to the model with the new marker and subscript 'old' to the model without it. Since the integrals of sensitivity and 'one minus specificity' over the (0, 1) interval can be seen as average sensitivity and 'one minus specificity', the IDI can be viewed as a difference between improvement in average sensitivity and any potential increase in average 'one minus specificity'. It can also be seen as an integrated difference in Youden's indices [28].

We show in the Appendix that the IDI can be estimated as follows:

$$\widehat{\mathrm{IDI}} = (\overline{\hat{p}}_{\mathrm{new,events}} - \overline{\hat{p}}_{\mathrm{old,events}}) - (\overline{\hat{p}}_{\mathrm{new,nonevents}} - \overline{\hat{p}}_{\mathrm{old,nonevents}}) \tag{13}$$

where $\overline{\hat{p}}_{\mathrm{new,events}}$ is the mean of the new model-based predicted probabilities of an event for those who develop events, $\overline{\hat{p}}_{\mathrm{old,events}}$ is the corresponding quantity based on the old model, $\overline{\hat{p}}_{\mathrm{new,nonevents}}$ is the mean of the new model-based predicted probabilities of an event for those who do not develop events and $\overline{\hat{p}}_{\mathrm{old,nonevents}}$ is the corresponding quantity based on the old model. Rearranging the terms in (13), we obtain

$$\widehat{\mathrm{IDI}} = (\overline{\hat{p}}_{\mathrm{new,events}} - \overline{\hat{p}}_{\mathrm{new,nonevents}}) - (\overline{\hat{p}}_{\mathrm{old,events}} - \overline{\hat{p}}_{\mathrm{old,nonevents}}) \tag{14}$$

which is a difference in discrimination slopes between the new and old models, as proposed by Yates [25, 26].

Since the actual events do not depend on the model, the standard deviation of $(\hat{p}_{\mathrm{new,events}} - \hat{p}_{\mathrm{old,events}})$ from equation (13) can be calculated as the standard error of paired differences of new and old model-based predicted probabilities across all event subjects, $\widehat{\mathrm{se}}_{\mathrm{events}}$. Denoting the corresponding estimator for non-events by $\widehat{\mathrm{se}}_{\mathrm{nonevents}}$ and assuming independence between events and non-events and their predicted probabilities, we obtain a simple asymptotic test for the null hypothesis of $\mathrm{IDI} = 0$:

$$z = \frac{\widehat{\mathrm{IDI}}}{\sqrt{(\widehat{\mathrm{se}}_{\mathrm{events}})^2 + (\widehat{\mathrm{se}}_{\mathrm{nonevents}})^2}} \tag{15}$$

Tests for means of dependent samples tend to have good asymptotic properties, so we expect the above statistic to be close to the standard normal for sufficiently large sample sizes. Individual components of the IDI assessing improvement separately for IS and integrated specificity can be tested using the approach of paired samples.

IDI vs improvement in AUC

The area under the sensitivity curve or, equivalently, the IS can be seen as an average sensitivity over the (0, 1) interval of possible cut-offs. The same is true for 'one minus specificity'. If we are willing to declare a new marker useful if it offers an improvement in sensitivity without an increase in 'one minus specificity', then we will be looking primarily at a difference in average (integrated) sensitivities. This is a regular or unweighted average. At the same time, a closer look at the formula for the AUC as an area under the plot of sensitivity *vs* 'one minus specificity' (see formula (A3) in the Appendix) reveals that it is also an average sensitivity, but this average is weighted by the derivative of specificity. Thus, the IDI and improvement in AUC are related in the sense that both can be seen as corrected average sensitivities—the IDI is corrected by the subtracted factor assessing the undesirable increase in 'one minus specificity', and the AUC by weighting the sensitivities of the two models of interest by the corresponding derivatives of specificities.

Positive and negative predictive values

It is important to note that the positive and negative predictive values (PPV and NPV) are two other important measures that complement sensitivity and specificity. It is natural to consider an improvement in PPV or NPV for a given cut-off or integrated over a range of cut-offs. However, this creates difficulties. An easy analog to the McNemar's test [27] for the improvement in sensitivity does not exist for PPV, since the denominators do not stay the same (numbers of subjects classified as events will vary from model to model). The usefulness of the PPV integrated over all possible cut-offs is substantially limited by the fact that large cut-offs lead to small denominators. An improvement of 5 per cent means something very different for denominators of 100 *vs* 10. This is not the case with IS, where the denominator remains fixed and small values of sensitivity have small effect on the overall measure. In conclusion, we recommend calculating PPV and NPV for a set of meaningful cut-offs with possible averaging over this meaningful set and using bootstrap methods for any formal testing.

APPLICATION

Effect of adding HDL cholesterol to coronary heart disease risk prediction models

The Framingham Heart Study started in 1948 [29] with the enrollment of the 'original' cohort of 5209 individuals. In 1971, 5124 participants (the offspring of the original cohort and their spouses) were enrolled into the Framingham Offspring Study. Of these, 3951 participants aged 30–74 attended the fourth Framingham Offspring cohort examination between 1987 and 1992. After excluding participants with prevalent CVD and missing standard risk factors including lipid levels, 3264 women and men remained eligible for this analysis. All participants gave written informed consent, and the study protocol was approved by the Institutional Review Board of the Boston Medical Center. Participants were followed for 10 years for the development of the first CHD event (myocardial infarction, angina pectoris, coronary insufficiency or CHD death). Cox proportional hazards models were employed with sex, diabetes and smoking as dichotomous and age, systolic blood pressure (SBP), total and HDL cholesterol (in one of the two models) as continuous predictors. The last three were standardized to facilitate comparisons of hazard ratios for the risk factors.

We evaluated the improvement in model performance introduced by the inclusion of HDL cholesterol using the indices described in previous sections. The increase in the AUC was evaluated and tested for significance using the test proposed by DeLong *et al.* [30]. IDI was estimated using formula (14), and test of significance was carried out as described by (15). We also examined and tested the improvement in the individual components (IS and IP) of IDI. Sensitivities and specificities at cut-off points of 6 and 20 per cent 10-year CHD risk were compared using the McNemar's test, and the change in PPV and NPV was assessed with bootstrap. Reclassification tables for subjects who do and do not develop events were constructed using <6, 6–20, >20 per cent 10-year CHD risk categories (based on the ATP III, cf. [5]). NRI and its components were estimated and tested for significance using the developments given in formulas (9)–(11). For simplicity of the presentation, we used 10-year predicted probabilities of a CHD event from the Cox model, with all performance measures calculated for a binary outcome ignoring time to event. Analyses for this paper were done using SAS software, version 9.1 (Copyright 2002–2003 SAS Institute Inc., Cary, NC, USA). All p-values given are two sided; the level of significance was set

DOI: 10.1002/sim

Table I. Cox regression coefficients for standard risk factors and HDL.

Variable	Parameter estimate	Standard error	p-value	Hazard ratio	95 per cent hazard ratio confidence limits	
Sex	−0.80875	0.17809	<0.0001	0.445	0.314	0.631
Diabetes	0.91487	0.21798	<0.0001	2.496	1.628	3.827
Smoking	0.31167	0.17316	0.0719	1.366	0.973	1.918
Age	0.05606	0.00922	<0.0001	1.058	1.039	1.077
St_SBP	0.16826	0.07889	0.0329	1.183	1.014	1.381
St_Total	0.15092	0.07239	0.0371	1.163	1.009	1.340
St_HDL	−0.42767	0.10322	<0.0001	0.652	0.533	0.798

St refers to standardized variables. Age is in years.

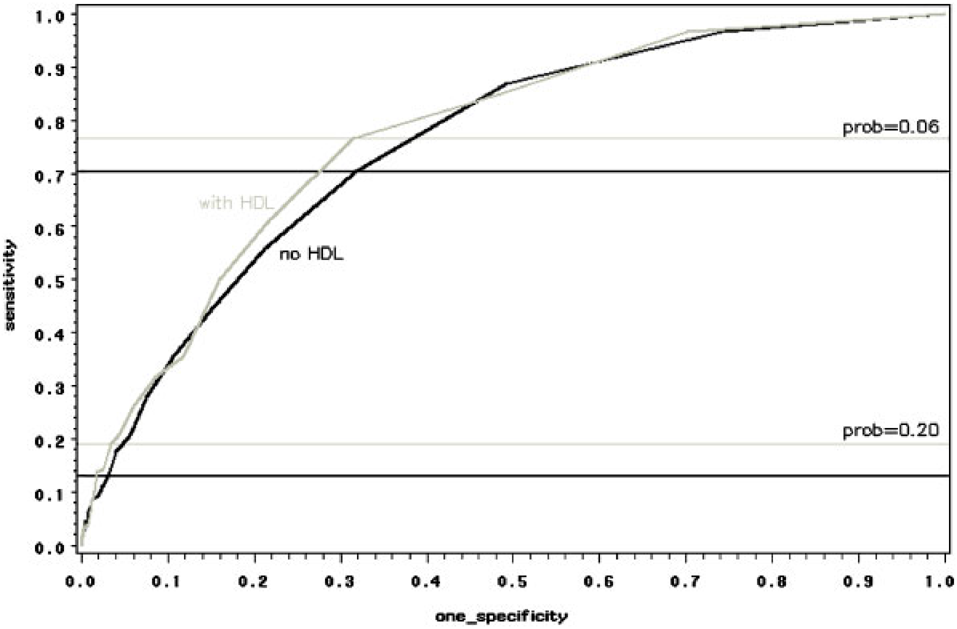

Figure 1. ROCs for models with and without HDL.

to 0.05. We note that this example is intended to illustrate concepts discussed in the paper rather than serve as a substantive analysis.

During the 10 years of follow-up, 183 individuals experienced a first CHD event. The hazard ratios and observed statistical significance levels for all predictors in the model including HDL are displayed in Table I. HDL was highly significant (HR = 0.65, p-value<0.001), and the AIC (Akaike Information Criterion [31]) decreased from 2779 to 2762 on addition of HDL. The ROC curves for the model without and with HDL are presented in Figure 1. The corresponding AUCs were 0.762 and 0.774, with the difference not statistically significant (p-value = 0.092). The sensitivities of models without and with HDL at the 6 and 20 per cent cut-offs were statistically significantly different (0.705 *vs* 0.765, p-value = 0.012 and 0.131 *vs* 0.191, p-value = 0.008,

DOI: 10.1002/sim

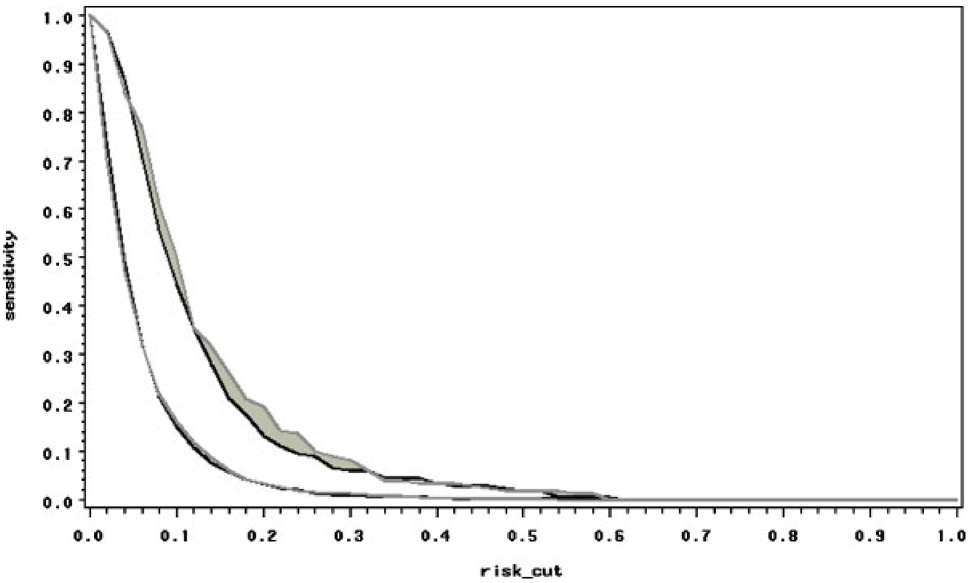

Figure 2. Sensitivity (top two lines) and 1-specificity as a function of risk cut-off. Models with (gray) and without HDL (black).

respectively). Specificities were not significantly different at either cut-off (without HDL: 0.682 *vs* 0.684, p-value = 0.682, with HDL: 0.968 *vs* 0.967, p-value = 0.508). The PPV increased from 0.116 to 0.126 for the 0.06 cut-off and from 0.197 to 0.254 for 0.20, with neither difference reaching statistical significance. No significant differences were noted in NPVs for either cut-off. IDI estimated at 0.009 was statistically significant (p-value = 0.008), mainly due to a 7 per cent increase in IS (0.120 *vs* 0.128 for the old *vs* new model, respectively; p-value = 0.022). The change in integrated 'one minus specificity' was not significant (0.0570 *vs* 0.0566, p-value = 0.283). These changes are depicted graphically in Figure 2.

Reclassifications for subjects with and without events are summarized in Table II. For 29 subjects who experienced CHD events, classification improved using the model with HDL, and for 7 people it became worse, with the net gain in reclassification proportion of 0.120, significantly greater than zero (p-value$<$0.001). The net gain in reclassification proportion for subjects who did not experience an event was not significant; 174 individuals were reclassified down and 173 were reclassified up (p-value = 0.957). The NRI was estimated at 0.121 and was highly significant (p-value$<$0.001).

HDL cholesterol is routinely used in CHD prediction models [3–5]. Both IDI and NRI suggest that including it in the prediction model results in significant improvement in performance. That conclusion could not have been drawn relying solely on the increase in AUC. The increase in IDI, albeit significant, was of small magnitude—0.009 on the absolute scale or 7 per cent relative increase. It can be interpreted as equivalent to the increase in average sensitivity given no changes in specificity. Based on the NRI and its components, we conclude that addition of HDL improved classification for a net of 12 per cent of individuals with events, with no net loss for non-events. Even though the NRI results look convincing, caution needs to be given to their interpretation, as it is dependent on the somewhat arbitrary choice of categories.

DOI: 10.1002/sim

Table II. Reclassification among people who experience a CHD event and those who do not experience a CHD event on follow-up.

Model without HDL	Model with HDL			
Frequency (Row per cent)	<6 per cent	6–20 per cent	>20 per cent	Total
Participants who experience a CHD Event				
<6 per cent	39 (72.22)	15 (27.78)	0 (0.00)	54
6–20 per cent	4 (3.81)	87 (82.86)	14 (13.33)	105
>20 per cent	0 (0.00)	3 (12.50)	21 (87.50)	24
Total	43	105	35	183
Participants who do not experience a CHD Event				
<6 per cent	1959 (93.24)	142 (6.76)	0 (0.00)	2101
6–20 per cent	148 (16.78)	703 (79.71)	31 (3.51)	882
>20 per cent	1 (1.02)	25 (25.51)	72 (73.47)	98
Total	2108	870	103	3081

DISCUSSION

In this paper, we proposed two new ways of assessing improvement in model performance accomplished by adding new markers—the net reclassification improvement (NRI) and integrated discrimination improvement (IDI). We derived sample estimators and simple asymptotic tests. Both measures are comprised of event and non-event components that can be examined separately. We also showed the relationship, between the new measures and how they relate to the improvement in AUC.

It is important to note that we did not address model calibration, another key measure of model performance. Calibration quantifies how closely the predicted probabilities of an event match the actual experience [7, 14, 26]. When evaluating the performance of a model after addition of a new marker, it is essential to check for improvement (or at least no adverse effect if other measures improve) in calibration, which can be quantified by, for example, the Hosmer–Lemeshow's chi-square or its modifications [32, 33].

The choice of the improvement metric should take into account the question to be answered—for example, if falling above or below a given cut-off is the primary basis for the choice of care, the cut-off specific improvement in sensitivity, specificity or the NRI might be the best choice. On the other hand, if no established risk cut-offs exist, the above might be of little use and the AUC or IDI might be preferred.

Additional characteristics of the IDI merit a mention. First, it is worth noting that since the discrimination slope suffers from the drawback of being dependent on model calibration, the same might also affect the IDI. Multiplying the predicted probabilities of event by a 'calibration factor', defined as a ratio of the observed event rate to the mean predicted probability of an event, will often substantially reduce this problem.

Second, since we expect to evaluate only new markers meant to improve model performance, the test given in (15) might be one sided. On the other hand, the IDI represents an averaged measure with reduced variability, and thus a more conservative significance level of 0.01 should be considered to avoid declaring too many markers useful. Another strategy would look at individual

DOI: 10.1002/sim

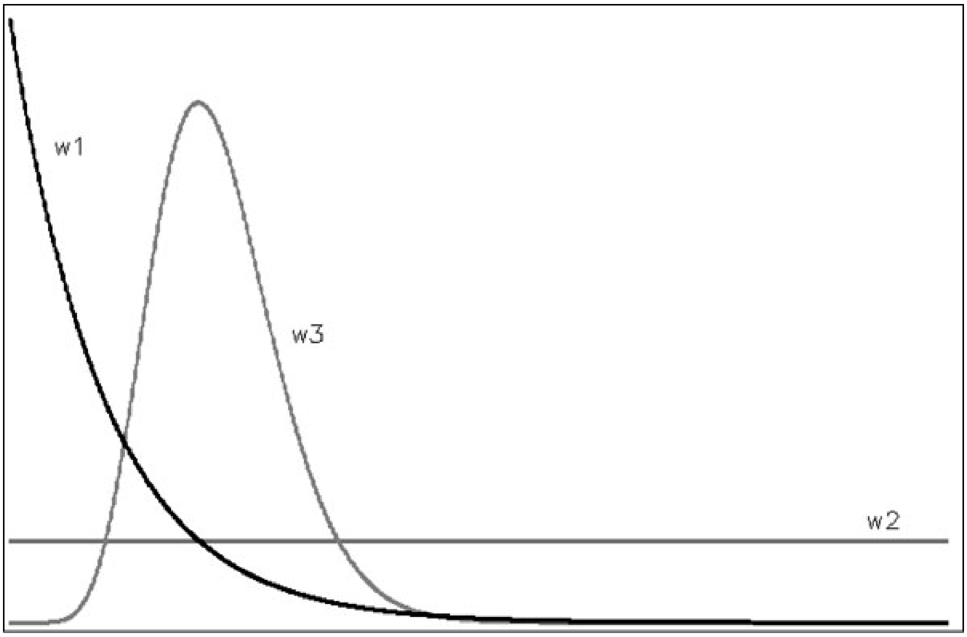

Figure 3. Weight functions. w1: AUC weight; w2: IS weight; w3: utility weight.

components of the IDI and perform only a test for difference in average sensitivities, provided no significant difference exists in average 'one minus specificities'.

We noted earlier that the AUC, IS and integrated 'one minus specificity' can be seen as weighted averages over the range of all cut-offs. The weight functions are depicted in Figure 3. Graph w1 corresponds to the derivative of specificity (or the pdf of predicted probabilities for non-events) used as a weight for sensitivity in the calculation of the AUC (cf. formula (A3) in the Appendix). Similarly, graph w2 is the weight used in the calculation of integrated sensitivity, IS. We note that the AUC weights more heavily values of sensitivity corresponding to small cut-offs, resulting in large sensitivities being weighted more heavily. This might explain why some apparent improvements in sensitivity for certain cut-offs do not translate into improvements in the AUC. The calculation of IS applies equal weight across all cut-offs or values of sensitivity. In many applications, some ranges of cut-off values are of more interest than others. A utility function of the type depicted by graph w3 in Figure 3 can be constructed to indicate which cut-offs are of particular importance. This utility function could be used as a weight function in the calculation of IS or integrated 'one minus specificity', leading to a weighted IDI.

Alternative weighting might also be considered to accommodate differences in the importance of sensitivity and specificity. In this context, the IDI and NRI could be presented as weighted differences of their components that correspond to improvements in sensitivity and specificity.

The discussion thus far did not draw any distinction between applications that do and do not consider time to event. Chambless and Diao [34] show that ignoring time to event in studies with long follow-up might lead to biased estimates of the AUC, sensitivity and specificity, and suggest ways to rectify the problem. The result of equation (14) can be generalized to the case of time-dependent sensitivity as described by Chambless' recursive definitions, and takes the form of a weighted average of the linear predictor in case of his alternative definitions.

We believe that NRI and IDI and their components offer useful insights into the process of assessing model improvement. We saw an example in which they suggested that HDL improves model performance—a finding consistent with current risk modeling practice that could be missed with sole reliance on the AUC. At the same time, we still believe that improvement in the AUC should remain the first criterion. But NRI and IDI should also be taken into consideration. For very large or very small differences in performance, all three quantities, improvement in AUC, IDI and NRI, should yield the same conclusions.

Our example also illustrates how increases in the AUC that would be considered small can lead to a substantial improvement in reclassification as quantified by the NRI and a more modest increase in the IDI. This might suggest reconsidering how to evaluate increases in the AUC—perhaps an increase of 0.01 might still be suggestive of a meaningful improvement. Interestingly, using the formulas presented in the Appendix, it can be deduced that assuming no change in specificity and a uniform increase in sensitivity of 0.01 at every cut-off point, both the AUC and IDI will go up by 0.01. This further illustrates how both can be seen as some forms of average sensitivity. But is an increase of 0.01 in average sensitivity satisfactory? This might depend on many factors. If specificity does not suffer and the additional burden and cost of obtaining the extra marker are low, it might be worth it to include it in the prediction models. If we deal with an expensive marker that is hard to obtain, we might want to resort to traditionally 'inexpensive' alternatives. As always, it is necessary to examine the clinical and public health implications of any decision.

APPENDIX

Mathematical developments

Below we present how the IDI can be estimated by a difference in discrimination slopes given by Yates [25, 26]. Let X represent the predicted probability of developing an event before time T and let D be the event indicator. If f is the probability density function (pdf) of X, for any cut-off point u, $0<u<1$, we can express sensitivity and 'one minus specificity' as

$$\text{Sensitivity}(u) = S(u) = P(X>u|D=1) = \int_u^1 f(x|D=1)\,\mathrm{d}x \tag{A1}$$

$$\text{OnemSpec}(u) = P(u) = P(X>u|D=0) = \int_u^1 f(x|D=0)\,\mathrm{d}x \tag{A2}$$

With the above notation, the AUC, which corresponds to the area under the parametric plot of $S(u)$ *vs* $P(u)$, can be expressed as (see Reference [16])

$$\text{AUC} = \int_0^1 S(u)\frac{\mathrm{d}}{\mathrm{d}u}P(u)\,\mathrm{d}u = \int_0^1 S(u)f(u|D=0)\,\mathrm{d}u = P(X_i>X_j|D_i=1, D_j=0) \tag{A3}$$

DOI: 10.1002/sim

Further define the integrated sensitivity and 'one minus specificity' as

$$\mathrm{IS}=\int_0^1 S(u)\,\mathrm{d}u=\int_0^1\int_u^1 f(x|D=1)\,\mathrm{d}x\,\mathrm{d}u \tag{A4}$$

$$\mathrm{IP}=\int_0^1 P(u)\,\mathrm{d}u=\int_0^1\int_u^1 f(x|D=0)\,\mathrm{d}x\,\mathrm{d}u \tag{A5}$$

Interchanging the order of integration in formulas (A4) and (A5), we get

$$\mathrm{IS}=\int_0^1\int_0^x f(x|D=1)\,\mathrm{d}u\,\mathrm{d}x=\int_0^1 xf(x|D=1)\,\mathrm{d}x=E(X|D=1) \tag{A6}$$

$$\mathrm{IP}=\int_0^1\int_0^x f(x|D=0)\,\mathrm{d}u\,\mathrm{d}x=\int_0^1 xf(x|D=0)\,\mathrm{d}x=E(X|D=0) \tag{A7}$$

The conditional expectations are estimated by sample means for events and non-events, that is $\overline{\hat{p}}_{\text{events}}=\sum_{i\text{ in events}}\hat{p}_i/\#$ events and $\overline{\hat{p}}_{\text{nonevents}}=\sum_{j\text{ in nonevents}}\hat{p}_j/\#$ nonevents, respectively. Hence, the integrated discrimination improvement (IDI) between the 'new' and 'old' models defined as

$$\mathrm{IDI}=(\mathrm{IS}_{\text{new}}-\mathrm{IS}_{\text{old}})-(\mathrm{IP}_{\text{new}}-\mathrm{IP}_{\text{old}}) \tag{A8}$$

can be estimated by

$$\widehat{\mathrm{IDI}}=(\overline{\hat{p}}_{\text{new,events}}-\overline{\hat{p}}_{\text{old,events}})-(\overline{\hat{p}}_{\text{new,nonevents}}-\overline{\hat{p}}_{\text{old,nonevents}}) \tag{A9}$$

or equivalently

$$\widehat{\mathrm{IDI}}=(\overline{\hat{p}}_{\text{new,events}}-\overline{\hat{p}}_{\text{new,nonevents}})-(\overline{\hat{p}}_{\text{old,events}}-\overline{\hat{p}}_{\text{old,nonevents}}) \tag{A10}$$

which is a difference of discrimination slopes.

ACKNOWLEDGEMENTS

This work was supported by the National Heart, Lung, and Blood Institute's Framingham Heart Study (contract N01-HC-25195 and 2K24 HL 04334). Dr D'Agostino Jr was partially supported by SEARCH for Diabetes in Youth funded by the Centers for Disease Control and Prevention (PA number 00097 and DP-05-069) and supported by the National Institute of Diabetes and Digestive and Kidney Diseases. Site Contract Number for the Coordinating Center (U01 DP000250).

REFERENCES

1. Kannel WB, McGee D. A general cardiovascular risk profile: the Framingham Study. *American Journal of Cardiology* 1976; **38**:46–51.
2. Anderson KM, Wilson PWF, Odell PM, Kannel WB. An updated coronary risk profile: a statement for health professionals. *Circulation* 1991; **83**:356–362.
3. Wilson PWF, D'Agostino RB, Levy D. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**:1837–1847.
4. D'Agostino RB, Russell MW, Huse DM *et al*. Primary and subsequent coronary risk appraisal: new results from the Framingham Heart Study. *American Heart Journal* 2000; **139**:272–281.

DOI: 10.1002/sim

5. 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). *Journal of the American Medical Association* 2001; **285**:2486–2497.
6. Conroy RM, Pyorala K, Fitzgerald AP *et al*. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal* 2003; **24**:987–1003.
7. D'Agostino RB, Grundy S, Sullivan LM, Wilson PWF. Validation of the Framingham Coronary Heart Disease Prediction Scores. *Journal of the American Medical Association* 2001; **286**(2):180–187.
8. Liu J, Hong Y, D'Agostino RB, Wu Z, Wang W, Sun J, Wilson PWF, Kannel WB, Zhao D. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese multi-provincial cohort study. *Journal of the American Medical Association* 2004; **291**:2591–2599.
9. Marrugat J, D'Agostino RB, Sullivan LM, Elosua R, Wilson PWF, Ordovas J, Solanas P, Cordón F, Ramos R, Sala J, Masiá R, Kannel WB. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. *Journal of Epidemiology and Community Health* 2003; **57**:634–638.
10. Ridker PM, Buring JE, Rifai N, Cook N. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. *Journal of the American Medical Association* 2007; **297**:611–619.
11. Lee DS, Evans JC, Robins SJ, Wilson PWF, Albano I, Fox CS, Wang TJ, Benjamin EJ, D'Agostino RB, Vasan RS. Gamma glutamyl transferase and metabolic syndrome, cardiovascular disease, and mortality risk: the Framingham Heart Study. *Arteriosclerosis Thrombosis Vascular and Biology* 2007; **27**:127–133.
12. Wang TJ, Gona P, Larson MG *et al*. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine* 2006; **355**:2631–2639.
13. D'Agostino RB. Risk prediction and finding new independent prognostic factors. *Journal of Hypertension* 2006; **24**:643–645.
14. D'Agostino RB, Griffith JL, Schmidt CH, Terrin N. Measures for evaluating model performance. *Proceedings of the Biometrics Section*, Alexandria, VA, U.S.A. American Statistical Association, Biometrics Section: Alexandria, VA. 1997; 253–258.
15. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**(1):29–36.
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. *Statistics in Medicine* 1996; **15**:361–387.
18. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine* 2004; **23**:2109–2123.
19. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* 2004; **159**:882–890.
20. Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. *Archives of Internal Medicine* 2005; **165**(21):2454–2456.
21. Ware JH. The limitations of risk factors as prognostic tools. *New England Journal of Medicine* 2006; **355**: 2615–2617.
22. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine* 2006; **145**:21–29.
23. Cook NR. Use and misuse of the receiver operating characteristics curve in risk prediction. *Circulation* 2007; **115**:928–935.
24. Pepe MS, Feng Z, Huang Y, Longton GM, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *UW Biostatistics Working Paper Series*, *#289.* 2006. Available at http://www.bepress.com/uwbiostat/paper289 (accessed 9 March 2007).
25. Yates JF. External correspondence: decomposition of the mean probability score. *Organizational Behavior and Human Performance* 1982; **30**:132–156.
26. Schmid CH, Griffith JL. Multivariable classification rules: calibration and discrimination. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: Chichester, U.K., 1998.
27. McNemar Q. Note on the sampling error of thee differences between correlated proportions or percentages. *Psychometrika* 1947; **12**:153–157.
28. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**:32–35.

DOI: 10.1002/sim

29. D’Agostino RB, Kannel WB. Epidemiological background and design: the Framingham Study. *Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions*, Washington, DC, U.S.A. American Statistical Association: Alexandria, VA, 1989.
30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.
31. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**:716–723.
32. D’Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of Statistics,* vol. 23. Elsevier Science B.V., 2004.
33. Hosmer Jr DW, Lemeshow S. *Applied Logistic Regression*. Wiley: New York, 1989.
34. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine* 2006; **25**:3474–3486.

DOI: 10.1002/sim