Disease Markers 26 (2009) 199–207
DOI 10.3233/DMA-2009-0631
IOS Press

Evaluating the prognostic value of new cardiovascular biomarkers

Angela M. Wood[a,*] and Philip Greenland[b]
[a]*Department of Public Health and Primary Care and Department of Biostatistics, University of Cambridge, Cambridge, UK*
[b]*Northwestern University Clinical and Translational Sciences Institute and Department of Preventive Medicine at Northwestern University Feinberg School of Medicine, Chicago, IL, USA*

Abstract. New predictors of cardiovascular outcomes are widely sought in research settings, and predictive tests are commonly recommended for routine use in cardiovascular clinical care. A number of multivariable scoring systems are in use around the world for assessment of a patient's risk. While such scoring systems are often recommended for clinical use in medical practice guidelines, their actual use in medical care falls short of recommendations. Limitations in the predictive capacity of existing predictive models are recognized, including lack of predictive accuracy, lack of ability to separate those who develop events from those who do not, and risks and costs of the testing modalities. Biomarker research is actively developing new testing strategies trying to improve upon current approaches, but it is often unclear how to assess the incremental prognostic information that a new test provides. In this report, we discuss the statistical approaches that can be used to evaluate additive predictive value of new tests. We also consider clinical research examples to put this information into a practical context.

Keywords: Statistics, biomarkers, prediction, cardiovascular risk

1. Introduction

The clinical approach to patients in cardiovascular disease treatment and prevention is commonly based on the clinician's ability to assess a patient's prognosis and target treatment intensity according to the severity of the patient's risk. Clinical practice guidelines in the United States [4,33] and Europe [13], as well as elsewhere in the world [50], recommend that clinicians utilize multivariable risk assessment approaches, especially in preventive cardiology, to select treatments for individual patients. Various approaches to cardiovascular risk assessment have been developed and recommended for general clinical use in the United States (e.g., Framingham Risk Score) and Europe (Systemic COronary Risk Evaluation (SCORE) model), and many other risk assessment models have been proposed [47]. Despite the prevalence of different risk assessment tools, they are not used as frequently as one might expect based on clinical practice guideline recommendations [3,14,30, 35,46], and there is relatively little evidence that current risk assessment methods actually lead to improved patient outcomes [47]. These sobering facts suggest that improved risk assessment methods are needed in cardiovascular prevention practice.

Given the limitations of current risk assessment tools and the associated problems with their application in practice, research directed toward finding new ways of predicting cardiovascular events is proceeding apace. These new methods cover a broad range of approaches including genomic tests [34], inflammation blood biomarkers such as C-reactive protein [9], thrombosis markers such as D-dimer [51], imaging tests such as coronary artery calcium measurement [10,19], vascular stiffness and vascular function tests [27], and biochemical profiles such as metabolomics [28] and other "omics" approaches. It has also been speculated that a combination of different types of markers, producing

[*] Corresponding author: Dr. Greenland, 750 North Lake Shore Drive, 11th Floor, Chicago, IL 60611, USA.

ISSN 0278-0240/09/$17.00 © 2009 – IOS Press and the authors. All rights reserved

a multimarker panel [17,53] would result in improved risk prediction and better targeting of risk reduction treatments.

With the profusion of new testing approaches, there is an urgent need to understand how to assess the additional predictive value of new markers, especially in regard to their additive value over standard and widely available predictors. For example, the Framingham Risk Score employs commonly measured variables such as patient age, gender, total blood cholesterol concentration, systolic blood pressure, HDL-cholesterol concentration, cigarette smoking history, and presence or absence of diabetes. In addition, many of these variables are amenable to interventions that are known to lower risk of cardiovascular disease, so a physician commonly targets treatment to the actual risk factors included in the risk assessment strategy. This may not be the case with imaging tests, vascular stiffness measurements, or genomic markers for which there are no known treatments at this time. Hence, unless a new predictive marker substantially improves risk prediction, or provides insight into a new way of treating risk, or replaces other less predictive markers, it may not have true clinical value.

In this paper, we discuss statistical approaches that are commonly used in evaluating whether a new test provides additive predictive information beyond that of a standard test battery. We also discuss ways in which clinicians can decide when a new test provides clinical utility and deserves consideration for routine use in clinical practice.

2. Measures of association

Measures of associations, such as hazard ratios, odds ratios and their corresponding confidence intervals, inform about the strength of the relationship between disease and the new marker, usually after adjustment for known markers. However, measures of association do not necessarily inform about the ability of the marker to predict disease [25,38,49,52]. Magnitudes of association measures will depend on the marker units (markers should be standardized so that the measure of association is given per standard deviation to avoid this limitation) and the significance of association generally depends on the sample size of the dataset. In addition, a strong association may suggest the marker is a good predictor for disease, but this may not translate to being clinically useful. This is because measures of association neither assess how accurate the predicted risks are, nor whether the predicted risks are sufficiently different for individuals with and without disease [38]. Later in this paper, we provide examples to illustrate the seeming paradox of a test being highly and consistently associated with a cardiovascular outcome but not additively predictive over standard risk assessment tools based on other statistical evaluations.

3. Measures of calibration

Assessing model calibration is an important tool in evaluating any risk prediction model. Measures of calibration assess how closely the predicted probabilities agree with the observed data. A sensible first step is to compare predicted versus observed risks, either in a graphical [7] or tabular display [25,39].

Calibration of a model is commonly summarized using the Hosmer-Lemeshow (H-L) test [23]. Individuals are categorized, typically into deciles [23], according to their predicted risk. The predicted risks are then compared against the observed risk in each category using squared differences. Significant differences between the observed and predicted risks produces small p-values (e.g.; $p < 0.05$), indicating poor calibration. A limitation is that the performance of the test is affected by how the categories are formed such that a contradictory test may result from an alternative grouping [7, 32]. It is also worth bearing in mind that the H-L test assesses the goodness-of-fit rather than puts a quantity on the accuracy of the predictions, and any deviation from the model assumptions (e.g.: linearity, proportional hazards) will affect the result. Thus, a poorly calibrated model may result from using the wrong functional form of the biomarker rather than indicating the absence of good predictors [24]. For this reason, and if the categories (deciles) are formed differently between two models, H-L test statistics will not always incrementally improve on addition of new markers.

4. Measures of explained variation

Measures of explained variation quantify the proportion of the variation in the observed disease outcome that can be explained by risk markers through a statistical model. In standard linear regression the most general definition is $\mathrm{R}^2 = 1$- (residual sum of squares)/(total sum of squares) and similar measures are available and in use for logistic regression [11,31]. However, measures of explained variation for survival models can be affected by the censoring nature of the event data, and thus there is no consensus on one measure, despite various proposals [43–45]. Values for R^2 and its equivalent measures typically range between 0 and 1: $R^2 = 1$ indicates that the fitted model explains all variability in the disease outcome while $R^2 = 0$ indicates no relationship between the disease and risk markers. Achieving a value close to 1 for survival models is extremely unlikely since it would mean that the model predicts the exact time point at which each participant fails (to whatever degree of precision the survival time is measured, e.g. to the day).

5. Measures of discrimination

Measures of discrimination quantify the separation in risk predictions between individuals with and without disease. The greater the separation in the risk predictions, the more likely the model will identify individuals at highest or lowest risk, and thus may be more clinically useful. The most popular measure of discrimination is the area under the ROC curve [AUROC], which is a function of sensitivity (the probability of a true positive test result) and the specificity (the probability of a true negative test result). This has been generalized for survival data by the C-index [20,21], with a definition of being the probability that for two randomly drawn patients, the person who has the event first has a higher probability of event. The c-index is estimated by examining all possible pairs of individuals under study for which the individual who has the shorter follow-up time fails. All possible pairs are classified as concordant (matching in rank according to the magnitude of the linear predictor and the order of failure), discordant (opposite in rank according to the magnitude of the linear predictor and the order of failure) or undecided (tied in either category). The overall measure is calculated as follows:

$$C = \frac{n_c + 0.5n_u}{n_c + n_d + n_u},$$

where n_c, n_d and n_u are the number of concordant, discordant and undecided pairs respectively. Note that censored individuals can only be compared to individuals known to fail before their censoring point and not to those who fail afterwards or to any other censored observations. Hence for datasets with large amounts of censoring, many pair-wise comparisons cannot be made which can result in bias [18].

Values for the AUROC and the c-index are fairly easily interpretable, being in the range of 0.5 to 1. A value of 0.5 indicates no ability, beyond that of chance, of the model to discriminate between participants in terms of risk; a value of 1 indicates perfect discrimination.

The AUROC describes how well a model can rank cases and non-cases, whereas the c-index compares the ranked survival times, incorporating information from censored individuals, but, neither measure is a function of the actual predicted probabilities. In some CVD prospective cohort studies, especially those examining younger women [39] the majority of individuals are at very low risk, with a small proportion being at high risk of disease. In such a circumstance, rank-based measures do not take this into account [16] so that 2 individuals who are at low risk (e.g.: 1.0% versus 1.1%) have the same impact on the AUROC and c-index as 2 individuals who are at moderate versus high risk (e.g.; 5% versus 20%). Thus, changes in the low ranks may lead to significant changes in the AUROC and c-index but could have very little clinical impact.

A recently proposed measure of discrimination for the survival model is the D-statistic [41]. The motivation for this measure is that it assesses the observed events across the spread of predictions and has an interpretation of being the log-hazard ratio comparing the upper-half predicted risk group versus the lower-half predicted risk group. Increasing values for D indicate greater separation between the observed risk of disease for participants predicted to be at high versus low risk. Royston and Sauerbrei [41] demonstrated the D-statistic to have many favorable properties, including interpretability, robustness to outliers and ability to deal with censoring. The D-statistic is easily computed by first transforming each participant's linear predictor from the fitted Cox model to transformed standard normal order rank statistics [41]. A second Cox regression on the rank statistic produces the coefficient D, interpreted as the log hazard ratio between individuals in the lower versus upper half predicted risk group.

It is worth noting that discrimination will generally depend on the range of predictor values available. For example, a study which includes individuals across a wide age range will typically have more ability to discriminate between individuals at high and low risk than a study with individuals from a narrow age range (and similarly for other strong predictors). This could lead to different studies having substantially different C-indices and D-statistics.

To assess the addition of a new marker to a model with established risk markers, it is more meaningful to investigate the change in measures for the R^2, AUROC, C-index and D-measure. These measures gen-

Table 1
Previously published measures of prognostic value for six progressive models for time to first coronary heart disease event, from all-male Malmo study ($n = 5983$, 653 CHD events over 23 years of follow-up). Modified from Figure 2 in Reference [49]

	R^2 (95% CI)	C-index (95% CI)	D-measure (95% CI)
age	0.00 (−0.00, 0.01)	0.60 (0.58, 0.62)	0.46 (0.32, 0.60)
age + systolic blood pressure	0.01 (0.01, 0.02)	0.65 (0.63, 0.67)	0.68 (0.55, 0.82)
Above + smoking	0.02 (0.01, 0.03)	0.67 (0.65, 0.69)	1.02 (0.89, 1.16)
Above + total cholesterol	0.03 (0.02, 0.04)	0.69 (0.67, 0.71)	1.10 (0.96, 1.23)
Above + fibrinogen	0.03 (0.02, 0.05)	0.69 (0.67, 0.71)	1.12 (0.99, 1.26)
Above + body mass index	0.04 (0.02, 0.05)	0.70 (0.68, 0.72)	1.16 (1.03, 1.30)

erally increase with each additional risk marker in the model, as illustrated in Table 1. Changes in the R^2, AUROC and c-index with addition of new markers to a standard predictive model are often very small and not easy to interpret. There is some controversy over the incremental change in the AUROC and c-index, when despite being small – still might be clinically meaningful [39]. Some believe that any improvement is important, whereas others may add a confidence interval on the incremental change and/or apply a formal test [40, 49].

6. Risk reclassification

Measures of explained variation or discrimination do not provide information about the actual risks that the models predict or about the proportion of participants who have low or high risk predictions. Reclassification methods however, as proposed by Cook, attempt to describe and measure the changes in predicted risk categories [6,7,39]. The primary purpose is to assess whether the addition of a new risk marker in a prediction model improves the classification of individuals into clinically relevant risk groups. The methods involve calculation of each participant's risk of CVD event at some pre-defined time point (typically 10 year risk) using models with and without a new biomarker of interest. Individuals are then classified into risk categories (e.g.: $<$ 5%, 5–10%, 10–20% and $>$ 20% risk of CVD event by 10 years) according to risk predictions from each model, and cross tabulated. The tables are appealing because they are visually interpretable, and simple calculations can be made such as the total percentage of individuals reclassified, or the percentage of people in an intermediate category who are then reclassified to a higher or lower risk category. The table may be grouped by individuals with and without events to distinguish individuals correctly and incorrectly reclassified [37].

Movement between categories for the two predictive models can be further summarised using measures recently proposed by Pencina et al. [37]. The Net Reclassification Improvement (NRI) summarises movement in the correct direction on average (i.e. events move up and non-events move down the risk categories). The NRI is a sum of two proportions: the proportion of individuals with events who move up or down the risk categories and the proportion of individuals without events who move up or down the risk categories. The NRI is difficult to interpret and it may be more useful to report the two components separately to be able to assess whether reclassification improves more for individuals with or without events [32,37,54]. An extension of the NRI is the clinical NRI, which summarises the movement between clinically relevant risk groups, e.g.: 10–20% risk groups to $>$ 20% say, is of interest if individuals above 20% risk are treated [5,40].

A limitation of reclassification tables and corresponding NRI measures is that they depend entirely on, and are affected by the choice of cut points, similarly to the Hosmer and Lemeshow test [37]. For example, suppose the risk categories are formed as follows: $<$ 5%, 5–10%, 10–20% and $>$ 20% risk of CVD event by 10 years. An individual who gets reclassified from a 19% predicted risk to 21% predicted risk will contribute to the NRI, unlike an individual who gets reclassified from 11% to 20% risk, despite the latter individual having a greater change. Using a different cut-off (e.g.: 15% or 19%) would lead to a different NRI. It would be highly clinically meaningful if a new biomarker reclassified people from low-risk (e.g.: $<$ 10%) to high risk (e.g.: $>$20%) but few examples have shown such large changes in the predicted risks. The Integrated Discrimination Improvement (IDI) which is a continuous version of the NRI summarises the improvement in 10-year risk prediction without categorisation into risk groups. Various relationships exist between the IDI and other known measures (such as difference in Yates slopes).

Another potential limitation of the NRI, and also the IDI, is that all directional reclassifications are assumed equivalent, such that movements between medium to high risk groups are treated with equal importance to movements between low and medium risk groups, despite the former being more clinically important. One solution would be to apply different weights to the different reclassifications, so that movements which are more clinically meaningful have more influence. Assigning such weights is likely to require further assumptions. It should also be noted that reclassification tables and the corresponding NRI, Clinical NRI and IDI measures exclude data from individuals censored within the risk period – thus often substantially reducing the available data from that used to estimate the risk prediction model. One solution is to look at the hazard ratio between those subjects who move into higher risk categories versus those moving into a lower risk category. Such an assessment can fully allow for the censored nature of the data within the risk period. However, all approaches censor events occurring after the predefined time point (e.g., 10 years), thus decreasing the number of available events and reducing event rates. It may be reasonable to consider risk reclassification at different time-points (e.g., 5, 10, 15 and 20 year risk) depending on the available length of follow-up.

7. Validation

An important feature of a risk prediction model is whether it is applicable to different patients [26,55]. The generalizability of a risk model depends heavily on the breadth of the population used to derive it. There may be little utility in deriving a risk model for CVD amongst males aged above 50 years only to try to predict the risk for a 30 year old female. Similarly, risk prediction models derived in Western populations may not be applicable to Asian populations [29]. These are extreme examples, and often one can only establish whether the risk prediction model works satisfactorily for similar patients to those from whose data it was derived [1,2]. Even for very similar patients, it cannot be assumed that the risk model will work well. Overoptimistic assessments of prognostic ability due to data-dependent methods used to derive the model are a well-known deficiency, and are exacerbated by small sample sizes. Further, the predictive ability of the model may be weak, and so even if the predictions for new individuals are unbiased, the model may still be unable to separate patients into clinically useful risk groups.

A common technique for establishing how well the risk model performs for new patients is "data-splitting" or "cross-validation" in the original data. The model is derived in one portion of the data and validated in the other. An important issue is how to split the data and although usually unbiased, it can lead to impression due to decreased sample sizes. Further, these validation methods do not address the wider issue of the generalizability of the model [42]. It is thus desirable to evaluate a model on appropriate data collected from another population [8,22].

8. Illustrative examples

There has been tremendous interest in the role of genomic testing as a way to enhance risk prediction in cardiovascular and other diseases. The futuristic concept of personalized medicine is predicated on the hope that genomic information will allow individualized risk assessments that will greatly exceed the predictive value of "population-based" approaches such as the Framingham Risk Score or the SCORE test. In cardiovascular medicine to date, this expectation has not been realized, as discussed recently [17]. One of the most consistently demonstrated associations with cardiovascular risk is the genetic variation at chromosome 9p21.3 [34]. Typically, this association yields a hazard ratio for prediction of cardiovascular events in the range of 1.1 to 1.4. Whether such an association produces predictive information beyond that of other readily available risk markers has not been fully examined. This additional predictive information was recently assessed in one large cohort in women who were initially free of any major chronic disease, prospectively followed over a median of 10.2 years for incident cardiovascular disease [34]. Polymorphism at rs10757274 was associated with an adjusted hazard ratio for incident cardiovascular disease of 1.25 (95% CI, 1.04 to 1.51) for the AG genotype and 1.32 (CI, 1.07 to 1.63) for the GG genotype. However, the addition of the genotype to a prediction model based on traditional risk factors, C-reactive protein, and family history of premature myocardial infarction had almost no effect on model discrimination as measured by the c-index (0.807 to 0.809) and did not improve the Net Reclassification Improvement score (-0.2%; $P = 0.59$) or the Integrated Discrimination Improvement score (0.0; $P = 0.18$). Whether this negative result will be replicated in more diverse cohorts is not known from this study, but this report demonstrates the paradoxical point discussed earlier in this paper that a consistent and statistically significant hazard ratio of association may fail to improve risk discrimination or net reclassification, such that it provides no meaningful improvement in clinical risk prediction beyond a standard battery.

Other tests, typically those with stronger degrees of association, are more likely to improve predictive discrimination [38]. The case of coronary artery calcium measurement provides an example of the potential for a relatively strong association to yield an improvement in predictive discrimination. In the Multi-Ethnic Study of Atherosclerosis (also called MESA), data on risk factors as well as measurements of coronary artery calcium score using rapid computed tomography were collected in a sample of 6722 previously healthy men and women from 4 major ethnic and racial groups in the United States [10]. Follow-up was for a median of 3.8 years, and there were 162 coronary events, of which 89 were major events (myocardial infarction or death from coronary heart disease). In comparison with participants with no coronary calcium, the adjusted hazard ratio of a coronary event was 7.73 (confidence intervals: 4.13–14.47) among participants with coronary calcium scores between 101 and 300 and 9.67 (CI, 5.20–17.98) among participants with scores above 300 ($P <$ 0.001 for both comparisons). As compared to the genomic study reported above in which AUROC did not appreciably change, the AUROC for the prediction of any coronary event was considerably higher when the calcium score was added to the standard risk factors (0.77 versus 0.82, $p <$ 0.001). This change in AUROC suggests a marked improvement in risk discrimination when the calcium score is added to standard risk predictive models. Similar results in improved C-statistics for calcium score added to Framingham Risk Scores have been reported in other studies [19].

As noted earlier, multimarker models have been speculated as a way to improve risk prediction. However, in practice, multimarker models have often not greatly improved risk discrimination when tested against standard risk models. Presumably, this occurs because many individual markers are highly collinear with other markers, and when combined as a multimarker panel, little new discrimination is seen. For example, Wang et al. [53] measured 10 promising biomarkers in 3209 participants attending a routine examination cycle of the Framingham Heart Study. These included blood tests of C-reactive protein, B-type natriuretic peptide, N-terminal pro–atrial natriuretic peptide, aldosterone, renin, fibrinogen, D-dimer, plasminogen-activator inhibitor type 1, and homocysteine; and the urinary albumin-to-creatinine ratio. During follow-up (median, 7.4 years), 207 participants died and 169 had a first major cardiovascular event. Adjusting for conventional risk factors, the biomarkers that most strongly predicted major cardiovascular events were B-type natriuretic peptide level (adjusted hazard ratio, 1.25 per 1 SD increment in the log values (95% confidence intervals (CI), 1.04–1.49) and the urinary albumin-to-creatinine ratio (1.20, 95% CI 1.02–1.41). Persons with “multimarker” scores (based on regression coefficients of significant biomarkers) in the highest quintile as compared with those with scores in the lowest two quintiles had elevated risks of death (adjusted hazard ratio, 4.08 (95% CI, 2.51–6.62); $P <$ 0.001) and major cardiovascular events (adjusted hazard ratio, 1.84 (95% CI, 1.11–3.05); $P =$ 0.02). However, the addition of multimarker scores to conventional risk factors resulted in relatively small increases in risk discrimination, as measured by the C-index (C-index for models of death were 0.75 (with age and sex as predictors), 0.79 (with age, sex, and multimarker score as predictors), 0.80 (with age, sex, and conventional risk factors as predictors), and 0.82 (with all predictors). Similar findings have been reported in other studies [12,36,48], but at least one recent study found a substantial increase in C-index upon addition of a novel multimarker panel to standard risk predictors [56]. This area requires further study to understand why some studies have found incremental risk prediction improvements [56] while many others have not.

9. Conclusions and clinical implications

Assessing utility of risk prediction models, and especially of new markers considered against previous standard predictive models, requires a broad-based statistical evaluation that exceeds statistical association tests alone. Tests of calibration, explained variation, discrimination, and reclassification can provide additional insights into whether new markers do, or do not, add predictive information to the standard markers. However, we have shown examples of statistically significant associations, often shown repeatedly to “predict” outcome, that have limited ability to improve discrimination or other statistical measures of predictive performance.

We have not considered here other important aspects of test performance that describe more advanced downstream effects of new tests in patient care. As discussed many years ago in regard to the clinical value of radiologic testing [15], there should be a hierarchical approach to assessing test efficacy that begins with technical efficacy of the test (such as ability of the test to be accurately measured and standardized in multiple laboratories) and proceeds to the highest level of efficacy – societal benefit. With increased emphasis on translational research, it should be emphasized that the most substantial and meaningful impact of any new test or procedure in medicine is an actual improvement in patient outcomes and a further benefit of this improvement on improved overall health of the population (such as can be measured by cost-benefit or cost-effectiveness analyses from a societal viewpoint). Between technical efficacy and societal benefit, there are additional steps of diagnostic or prognostic accuracy (such as a measure of association); diagnostic thinking efficacy (change in the physician's understanding of the patient's risk based on the added value of the new test); therapeutic efficacy (e.g., number of times a new test actually resulted in changes in patient treatments); and patient outcome efficacy (e.g., proportion of patients in whom cardiovascular events were prevented due to the improved selection of treatments following better risk prediction). We recognize, with great humility, that most current efforts in biomarker discovery and validation focus on the earliest levels of this hierarchical model when conducting statistical and technical studies of test efficacy. Only when we can demonstrate the important downstream improvements in therapeutic changes and in actual patient or societal outcomes can we be confident that new markers have actually created the reality of "personalized and predictive medicine." This remains a hope at this point and not yet a reality, beyond the demonstrated value of traditional risk factors for targeting patients for known risk reduction strategies such as cholesterol lowering and anti-hypertensive treatments [4,33].

References

[1] D.G. Altman and P. Royston, What do we mean by validating a prognostic model? *Stat Med* **19** (2000), 453–473.

[2] P. Brindle, J. Emberson, F. Lampe, M. Walker, P. Whincup and S. Ebrahim, Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study, *BMJ* **327** (2003), 1267–1270.

[3] M.D. Cabana, C.S. Rand, N.R. Powe, A.W. Wu, M.H. Wilson, P.A. Abboud and H.R. Rubin, Why don't physicians follow clinical practice guidelines? A framework for improvement, *JAMA* **282** (1999), 1458–1465.

[4] A.V. Chobanian, G.L. Bakris, H.R. Black, W.C. Cushman, L.A. Green, J.L. Izzo Jr., D.W. Jones, B.J. Materson, S. Oparil, J.T. Wright Jr. and E.J. Roccella, National Heart, Lung, and Blood Institute Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure; National High Blood Pressure Education Program Coordinating Committee. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report, *JAMA* **289** (2003), 2560–2572.

[5] N. Cook, Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M.J. Pencina et al., Statistics in Medicine, *Stat Med* **27** (2008), 191–195.

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

[7] N.R. Cook, J.E. Buring and P.M. Ridker, The effect of including C-reactive protein in cardiovascular risk prediction models for women, *Ann Intern Med* **145** (2006), 21–29.

[8] R.B. D'Agostino Sr., S. Grundy, L.M. Sullivan and P. Wilson, CHD Risk Prediction Group, Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation, *JAMA* **286** (2001), 180–187.

[9] J. Danesh, J.G. Wheeler, G.M. Hirschfield, S. Eda, G. Eiriksdottir, A. Rumley, G.D. Lowe, M.B. Pepys and V. Gudnason, C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease, *N Engl J Med* **350** (2004), 1387–1397.

[10] R. Detrano, A.D. Guerci, J.J. Carr, D.E. Bild, G. Burke, A.R. Folsom, K. Liu, S. Shea, M. Szklo, D.A. Bluemke, D.H. O'Leary, R. Tracy, K. Watson, N.D. Wong and R.A. Kronmal, Coronary calcium as a predictor of coronary events in four racial or ethnic groups, *N Engl J Med* **358** (2008), 1336–1345.

[11] B. Efron, The efficiency of logistic regression compared to normal discriminant analysis, *J Am Stat Assoc* **70** (1975), 892–898.

[12] A.R. Folsom, L.E. Chambless, C.M. Ballantyne, J. Coresh, G. Heiss, K.K. Wu, E. Boerwinkle, T.H. Mosley Jr., P. Sorlie, G. Diao and A.R. Sharrett, An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study, *Arch Intern Med* **166** (2006), 1368–1373.

[13] Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice, European guidelines on cardiovascular disease prevention in clinical practice: executive summary, *Eur Heart J* **28** (2007), 2375–2414.

[14] J.P. Frolkis, S.J. Zyzanski, J.M. Schwartz and P.S. Suhan, Physician noncompliance with the 1993 National Cholesterol Education Program (NCEP-ATPII) guidelines, *Circulation* **98** (1998), 851–855.

[15] D.G. Fryback and J.R. Thornbury, The efficacy of diagnostic imaging, *Med Decis Making* **11** (1991), 88–94.

[16] T.A. Gerds, T. Cai and M. Schumacher, The performance of risk prediction models, *Biometrical Journal* **50** (2008), 457–479.

[17] R.E. Gerszten and T.J. Wang, The search for new cardiovascular biomarkers, *Nature* **451** (2008), 949–952.

[18] M. Gonen and G. Heller, Concordance probability and discriminatory power in proportional hazards regression, *Biometrika* **92** (2005), 965–970.

[19] P. Greenland, L. LaBree, S.P. Azen, T.M. Doherty and R.C. Detrano, Coronary artery calcium score combined with Fram-

ingham score for risk prediction in asymptomatic individuals, *JAMA* **291** (2004), 210–215.

[20] F.E. Harrell Jr., R.M. Califf, D.B. Pryor, K.L. Lee and R.A. Rosati, Evaluating the yield of medical tests, *JAMA* **247** (1982), 2543–2546.

[21] F.E. Harrell Jr., K.L. Lee and D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Stat Med* **15** (1996), 361–387.

[22] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson and P. Brindle, Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study, *Heart* **94** (2008), 34–29.

[23] D.W. Hosmer and S. Lemeshow, Applied Logistic Regression, New York: Wiley, 1989, Section 5.2.2.

[24] D.W. Hosmer, T. Hosmer, S. Le Cessie and S. Lemeshow, A comparison of goodness-of-fit tests for the logistic regression model, *Stat Med* **16** (1997), 965–980.

[25] H. Janes, M.S. Pepe and W. Gu, Assessing the value of risk predictions by using risk stratification tables, *Ann Intern Med* **149** (2008), 751–760.

[26] A.C. Justice, K.E. Covinsky and J.A. Berlin, Assessing the Generalizability of Prognostic Information, *Ann Intern Med* **130** (1999), 515–524.

[27] S. Laurent, J. Cockcroft, L. Van Bortel, P. Boutouyrie, C. Giannattasio, D. Hayoz, B. Pannier, C. Vlachopoulos, I. Wilkinson and H. Struijker-Boudier, European Network for Non-invasive Investigation of Large Arteries, Expert consensus document on arterial stiffness: methodological issues and clinical applications, *Eur Heart J* **27** (2006), 2588–2605.

[28] G.D. Lewis, A. Asnani and R.E. Gerszten, Application of metabolomics to cardiovascular biomarker and pathway discovery, *J Am Coll Cardiol* **52** (2008), 117–123.

[29] J. Liu, Y. Hong, R.B. D'Agostino Sr., Z. Wu, W. Wang, J. Sun, P.W. Wilson, W.B. Kannel and D. Zhao, Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study, *JAMA* **291** (2004), 2591–2599.

[30] P. McBride, H.G. Schrott, M.B. Plane, G. Underbakke and R.L. Brown, Primary care practice adherence to National Cholesterol Education Program guidelines for patients with coronary heart disease, *Arch Intern Med* **158** (1998), 1238–1244.

[31] D. McFadden, Conditional logit analysis of qualitative choice behavior, in: *Frontiers in Economics*, P. Zarembka, ed., NY: Academic Press, 1974.

[32] K. McGeechan, P. Macaskill, L. Irwig, G. Liew and T.Y. Wong, Assessing New Biomarkers and Predictive Models for Use in Clinical Practice: A Clinician's Guide, *Arch Intern Med* **168** (2008), 2304–2310.

[33] National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III), 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) final report, *Circulation* **106** (2002), 3143–421.

[34] N.P. Paynter, D.I. Chasman, J.E. Buring, D. Shiffman, N.R. Cook and P.M. Ridker, Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3, *Ann Intern Med* **150** (2009), 65–72.

[35] T.A. Pearson, I. Laurora, H. Chu and S. Kafonek, The lipid treatment assessment project (L-TAP): a multicenter survey to evaluate the percentages of dyslipidemic patients receiving lipid-lowering therapy and achieving low-density lipoprotein cholesterol goals, *Arch Intern Med* **160** (2000), 459–467.

[36] A. Peer, G. Falkensammer, H. Alber, A. Kroiss, A. Griesmacher, H. Ulmer, O. Pachinger and J. Mair, Limited utilities of N-terminal pro B-type natriuretic peptide and other newer risk markers compared with traditional risk factors for prediction of significant angiographic lesions in stable coronary artery disease, *Heart* **95** (2009), 297–303.

[37] M.J. Pencina, A.R. D' Sr., A.R. D' Jr. and R.S. Vasan, Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond, *Stat Med* **27** (2007), 157–172.

[38] M.S. Pepe, H. Janes, G. Longton, W. Leisenring and P. Newcomb, Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker, *Am J Epidemiol* **159** (2004), 882–890.

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

[40] P.M. Ridker, N.P. Paynter, N. Rifai, J.M. Gaziano and N.R. Cook, C-Reactive Protein and Parental History Improve Global Cardiovascular Risk Prediction, The Reynolds Risk Score for Men, *Circulation* **118** (2008), 2243–2251.

[41] P. Royston and W. Sauerbrei, A new measure of prognostic separation in survival data, *Stat Med* **23** (2004), 723–748.

[42] P. Royston, M.K. Parmar and R. Sylvester, Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer, *Stat Med* **23** (2004), 907–926.

[43] M. Schemper, The explained variation in proportional hazards regression, *Biometrika* **77** (1990), 216–218.

[44] M. Schemper and R. Henderson, Predictive accuracy and explained variation in Cox regression *Biometrics* **56** (2000), 249–255.

[45] M. Schemper and J. Stare, Explained variation in survival analysis, *Stat Med* **15** (1996), 1999–2012.

[46] S. Sheridan, M. Pignone and C. Mulrow, Framingham-based tools to calculate the global risk of coronary heart disease: a systematic review of tools for clinicians, *J Gen Intern Med* **18** (2003), 1039–1052.

[47] S.L. Sheridan and E. Crespo, Does the routine use of global coronary heart disease risk scores translate into clinical benefits or harms? A systematic review of the literature, *BMC Health Serv Res* **8** (2008), 60.

[48] M.G. Shlipak, L.F. Fried, M. Cushman, T.A. Manolio, D. Peterson C. Stehman-Breen, A. Bleyer, A. Newman, D. Siscovick and B. Psaty, Cardiovascular mortality risk in chronic kidney disease: comparison of traditional and novel risk factors, *JAMA* **293** (2005), 1737–1745.

[49] The Fibrinogen Studies Collaboration, Measures to assess the prognostic ability of the stratified Cox proportional hazards model, *Stat Med* **28** (2009), 389–411.

[50] A. Tonkin, P. Barter, J. Best, A. Boyden, J. Furler, K. Hossack, D. Sullivan, P. Thompson, M. Vale, C. Cooper, M. Robinson and E. Clune, National Heart Foundation of Australia; Cardiac Society of Australia and New Zealand. National Heart Foundation of Australia and the Cardiac Society of Australia and New Zealand: position statement on lipid management–2005, *Heart Lung Circ* **14** (2005), 275–291.

[51] H. Vidula, L. Tian, K. Liu, M.H. Criqui, L. Ferrucci, W.H. Pearce, P. Greenland, D. Green, J. Tan, D.B. Garside, J. Guralnik, P.M. Ridker, N. Rifai and M.M. McDermott, Biomark-

ers of inflammation and thrombosis as predictors of near-term mortality in patients with peripheral arterial disease: a cohort study, *Ann Intern Med* **148** (2008), 85–93.

[52] N.J. Wald, A.K. Hackshaw and C.D.Frost, When can a risk factor be used as a worthwhile screening test? *Br Med J* **319** (1999), 1562–1565.

[53] T.J. Wang, P. Gona, M.G. Larson, G.H. Tofler, D. Levy, C. Newton-Cheh, P.F. Jacques, N. Rifai, J. Selhub, S.J. Robins, E.J. Benjamin, R.B. D'Agostino and R.S. Vasan, Multiple biomarkers for the prediction of first major cardiovascular events and death, *N Engl J Med* **355** (2006), 2631–2639.

[54] P.W.R. Wilson, M. Pencina, P. Jacques, J. Selhub, R. D'Agostino Sr. and C.J. O'Donnell, C-Reactive Protein and Reclassification of Cardiovascular Risk in the Framingham Heart Study, *Circulation: Cardiovascular Quality and Outcomes* **1** (2008), 92–97.

[55] J.C. Wyatt and D.G. Altman, Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ* **311** (1995), 1539–1541.

[56] B. Zethelius, L. Berglund, J. Sundström, E. Ingelsson, S. Basu, A. Larsson, P. Venge and J. Arnlöv, Use of multiple biomarkers to improve the prediction of death from cardiovascular causes, *N Engl J Med* **358**(20) (2008), 2107–2116.