Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2011;40:1094–1105
doi:10.1093/ije/dyr013

Use of reclassification for assessment of improved prediction: an empirical evaluation

Ioanna Tzoulaki,[1,2] George Liberopoulos[2] and John P A Ioannidis[2,3,4,5]*

[1]Department of Epidemiology and Biostatistics, Imperial College of Medicine, London, UK, [2]Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, [3]Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece, [4]Department of Medicine, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, and Tufts University School of Medicine, Boston, MA, USA and [5]Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA, USA

*Corresponding author. Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, CA 94305, USA. E-mail: jioannid@stanford.edu

Accepted 11 January 2011

Background An increasing number of studies evaluate the ability of predictors to change risk stratification and alter medical decisions, i.e. reclassification performance. We examined the reported design and analysis of recent studies of reclassification and the robustness of their claims for improved reclassification.

Methods Two independent investigators searched PubMed and citations to the article that introduced the currently most popular reclassification metric (net reclassification index, NRI) to identify studies performing reclassification analysis (January 2006–January 2010). We focused on articles that included any analyses comparing the performance of a baseline predictive model vs the baseline model plus some additional predictor for a prospectively assessed outcome. We recorded information on the baseline model used, outcomes assessed, choice of risk thresholds and features of reclassification analyses.

Results Of 58 baseline models used in 51 eligible papers, only 14 (24%) were previously described, used as described and had same outcomes as originally intended. Calibration was examined in 53% of the studies. Sixteen studies (31%) provided a reference for the choice of risk thresholds and only six used the previously proposed categories or justified the use of alternative thresholds. Only 14 studies (27%) stated that the chosen risk thresholds had different therapeutic intervention implications. NRI was calculated in 38 studies and was smaller in studies with adequately referenced or justified risk thresholds vs others ($P < 0.0001$).

Conclusions Reclassification studies would benefit from more rigorous methodological standards; otherwise claims for improved reclassification may remain spurious.

Keywords Reclassification, predictive models, risk stratification

Introduction

Assessment of risk for disease development or progression is fundamental to decision making in medicine and public health.[1] For example, preventive therapy for coronary heart disease (CHD) development can be guided by the Framingham risk score (FRS), a multivariable risk prediction model incorporating information on individual's age, gender, serum

cholesterol levels, systolic blood pressure, diabetes status and smoking.[2] Patients beyond a certain level of risk should be offered preventive treatment, whereas patients below a certain level of risk should not. Similar examples exist for a wide range of disease outcomes, ranging from intensive care unit outcomes, cancer, fracture risk and so forth.[3–8] Moreover, with advances in research, novel markers of risk, which can potentially improve prediction over and above established risk prediction models, are constantly being proposed. Many envisage that this information will eventually lead to 'personalized medicine' of patient-specific treatments.[9] However, demonstrating that a new candidate predictor can improve prediction beyond traditional risk factors is a demanding task and requires careful attention to study design and analysis to avoid unjustified or premature claims.[10]

The literature of predictive modelling has recently moved beyond the evaluation of discrimination, commonly assessed with the ROC curve, to the ability to change risk stratification, the so-called reclassification.[11–13] In contrast to area under the curve (AUC) analysis, reclassification does not examine just the ability to improve the accuracy of prediction in general; instead, it may provide clinically meaningful improvements in risk prediction. Reclassification measures focus specifically on the ability to classify people more appropriately in risk categories that have different implications for treatment. This could lead to better treatment choices and thus better outcomes. However, there are many prerequisites for this goal to be achieved:[14] effective treatments must be available; indications for treatment must vary per level of risk; the thresholds of risk that dictate different treatment must be well defined; and baseline models (to which addition of a new predictor is contemplated) must be well defined and standardized with widely accepted included traditional predictors and clinical outcomes. Otherwise, reclassification studies may become susceptible to diverse biases with subjective definitions, analyses and reporting thereof.

Here, we assessed empirically a systematic sample of recent studies that evaluated the reclassification ability of various candidate predictors. We aimed to examine the reported design and analysis of these studies and the robustness of their claims for improved reclassification.

Methods

Eligibility criteria and selection of studies

We aimed to assemble a sample of recent studies that examined one or more candidate risk factors' ability to reclassify individuals into different risk categories compared with a predictive model that does not include these markers. For consistency, we aimed to focus on studies that addressed the incremental predictive ability of one or more candidate predictors. Thus, we did not consider studies that compared completely different models that did not differ simply on the presence or not of specific additional markers.

We used two approaches to search for eligible articles. We performed a PubMed search using the following algorithm '(reclassif* [tw] OR re-classif* [tw]]) AND (predict* OR progn*)' limited to human studies for the period January 2006– January 2010, because the concept of reclassification was uncommon before then; and also searched until January 2010 for citations to the most highly cited methods paper on re-classification that introduced and popularized the net reclassification index (NRI).[11] We first perused the title and abstract of each of these citations. Potentially eligible articles were retrieved for perusal in full text.

Studies were eligible if they had primary data (not review, guideline, letter, editorial, etc.) on real patients (not decision analyses with simulated data) and regardless of field, type of predictive model and outcome. We included studies regardless of whether they gave exact quantitative results of reclassification metrics, provided that they stated that they did such reclassification analysis. We considered studies with longitudinal follow-up (prospective or retrospective), and excluded cross-sectional studies (they pertain to diagnosis rather than prognosis). When an article considered separately two or more predictors for their incremental predictive ability, information was considered separately for each of the examined additional predictors. When several additional predictors were considered together in all analyses, we did not separate them. When an article considered more than one outcome, we evaluated these separately.

Data extraction

For each eligible article, we recorded the first author, journal, year of publication, number of subjects in the study, additional predictor(s), baseline predictive model, outcome(s) assessed, population evaluated and features of reclassification analyses.

For the baseline model, we extracted information on how each model was modelled, i.e. whether a risk score was calculated based on previously published coefficients or whether a set of variables was used to develop/reconstruct the model, and on whether it was already described in the literature and referenced; alluded to have been described already in the literature, but not referenced; unclear whether it had been already described in the literature; or a model that the authors claimed to build for the first time—and, if so, how this was justified. Whenever a model was already described in the literature, we examined whether the authors used it as previously described or deviated somehow with inclusion of other variables, exclusion of standard variables, substitution of standard variables with other variables, different definitions or different modelling of standard variables.

We also extracted the risk category thresholds and whether their choice was referenced (and, if so, whether the reference indeed proposed that specific categorization or not) or otherwise justified and how; whether more than one categorization was investigated (and, if so, whether results were given for all categorizations vs selectively, and whether only the results with best reclassification were shown); and whether it was stated that the risk categories chosen carried also different therapeutic/intervention implications or not. We also examined whether the statement of different therapeutic/intervention implications was supported by a reference that indeed proved this point or other arguments, and the specific intervention(s) that would differ based on the risk categories.

For each additional candidate predictor, we recorded how it was entered in the analyses, noting in particular whether different options were examined and then a best-fitting option was used in the evaluation of incremental predictive ability.

For reclassification features, we recorded whether authors reported the predicted risk for patients categorized in each risk category for the baseline model and for the model including also the additional predictor(s); the number or percentage of patients changing risk categories for each type of change; and whether participants who developed the disease moved into a higher risk category and those who did not develop the disease to a lower risk category (i.e. risk prediction moved in the correct direction). We also noted whether authors used the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI).[11] The NRI is calculated by summing across risk categories the proportion of participants whose estimated risk shifts in the correct direction minus the proportion of those who shift into the wrong direction among case patients vs control participants, that is, those who did or did not develop the disease during follow-up. Participants who had an event are thought to shift in the correct direction if they move into a higher estimated risk category and to shift in the wrong direction if they move into a lower estimated risk category. The opposite applies for participants who did not have an event. The IDI does not use discrete risk categories, but calculates the difference in Yates, or discrimination, slopes between two models. Because NRI and IDI depend on model calibration (how closely predicted estimates of absolute risk agree with actual outcomes) we also recorded whether authors examined calibration of models (and if so, how) and whether calibration results were presented. We also noted whether NRI was calculated based on time-to-event data (Cox regression) and if authors accounted for incomplete follow-up on NRI calculation. We extracted NRI and IDI values and 95% confidence intervals (CIs) from studies that performed such analyses. We calculated the CI when this was not given by the authors but other data to calculate it were available. Whenever a study examined subgroups, we focused on the whole population unless only data per subgroup were provided; in those cases, we extracted data for each subgroup separately. We focused on main analyses rather than any additional sensitivity analyses. Finally, we noted whether authors claimed improved reclassification when interpreting their results.

We aimed to examine whether studies with different methodological characteristics have different NRI values on average. We compared the NRI in studies that examined baseline models that had been previously described, used as described, and had the same outcomes as originally intended (adequate baseline model) vs other studies; and in studies that used the same risk thresholds as the originally described in the references or justified the use of alternative thresholds (adequate threshold) vs other studies. For each subgroup, NRI values were combined with random effects so as to obtain a summary NRI for each subgroup. Random effects aim to estimate an average of the population of NRI values, but it allows that the actual true NRI values of single studies may differ, as is expected for studies of different diseases, outcomes, predictors, and study-specific biases. Random effects were preferred since we anticipated that NRI estimates from such diverse fields and predictors would have unavoidable heterogeneity.[15,16] The two summary NRI values were compared with a *Q*-test in order to examine whether studies with different methodological characteristics have on average different NRI values. We also performed these analyses limited only to studies that had cardiovascular outcomes, so as to have more homogeneous sets of studies to compare. The other disease groups are too sparse in numbers to perform any meaningful analyses.

Analyses were performed in Comprehensive Meta-analysis version 2.2.050. *P*-values are two-tailed.

Results

Eligible studies

A total of 75 items retrieved from PubMed and 63 retrieved from the citations of Pencina *et al.*[11] were considered potentially eligible and perused in full text and of those 34 and 40 articles, respectively, were eligible (Supplementary Figures 1 and 2). Twenty-three articles were identified from both search strategies, thus 51 articles were included.[17–67] The κ- coefficient for eligibility between the two independent investigators on initial screening was 0.80. Main study characteristics appear in Supplementary Table 1. As shown, most studies pertained to cardiovascular outcomes or mortality and had been published in major general medical journals or cardiovascular journals. The median sample size was 3441 [Interquartile range (IQR), 1406–10 724].

Baseline model

Fifty-eight baseline models were examined in the 51 eligible papers; of those only 14 (24%) were previously described, used as described and had the same outcomes as originally intended (Figure 1). These models were FRS[2,68] for CHD ($n=5$), Framingham offspring study[69] and Cambridge scores[70] for type 2 diabetes ($n=2$ and 1, respectively), MELD score[8] for mortality in patients with end-stage liver disease ($n=2$), the 'six simple variable'[71] model for predicting outcome after acute stroke ($n=1$), blood pressure[72] for development of hypertension ($n=1$), SCORE for CVD mortality[73] ($n=1$), and a nomogram predicting fracture risk[74] ($n=1$).

Additional predictors

Forty-five different predictors were assessed (Supplementary Table 3). Most common predictors included genetic variants and C-reactive protein (eight studies each). In 17 analyses more than one options of modelling the additional predictor were stated to have been examined, and in 12 results were presented for only one option [due to best fit ($n=2$), similar results ($n=1$) or unjustified ($n=9$)].

Risk categories

Table 1 lists the different thresholds used in each analysis for outcomes examined in more than one study (thresholds for remaining outcomes are presented in Supplementary Table 4). Different thresholds, which were not referenced or otherwise justified, were used even when studies examined the same outcome for the same time period. For example, among 10 studies that examined CHD in general populations, five used <5/5–10/10–20/>20% risk thresholds, two used <10/10–20/>20%, one used <5/5–10/10–15/>15%, one used <6/6–20/>20% and one used <10/10–25/25–30/>30% for percentage of patients with CHD event at 10 years. Seven of these 10 studies used the FRS (or FRS covariates) as the baseline model that is associated with use of <10/10–20/>20% risk categories. Inadequate use of thresholds was also observed for other endpoints including CVD, CVD mortality and type 2 diabetes.

Six studies stated that they evaluated two different risk classification thresholds for the same outcome and five presented results on both. One study did not state the adopted risk thresholds, simply stating that participants 'were reclassified into adequate risk categories'.[63]

Only 16 studies (31%) provided a reference for the choice of risk thresholds and 14 studies (27%) stated that the chosen risk thresholds had different intervention implications (Figure 1). The most common intervention associated with different risk thresholds was lipid-lowering therapy (eight studies), whereas others

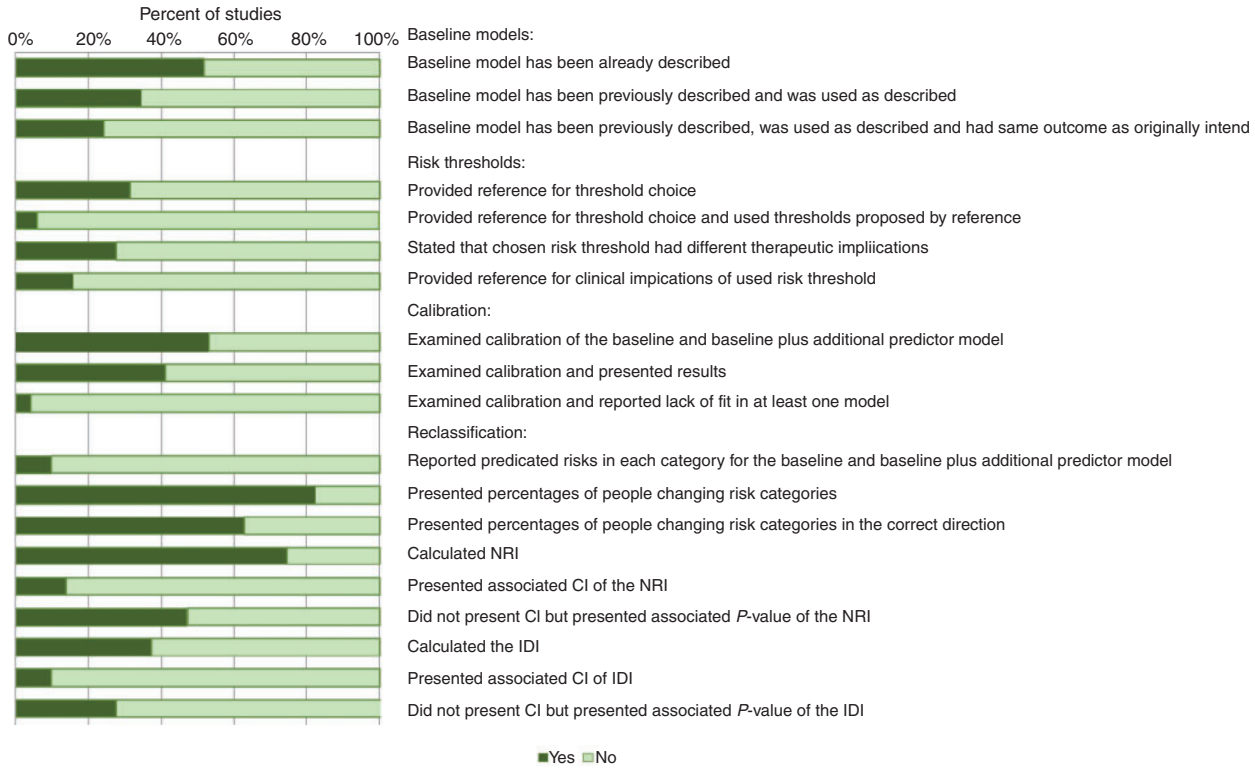

Figure 1 Characteristics of the 51 eligible studies in relation to baseline model used, risk threshold use and definition for reclassification analyses, examination of model calibration, and use of reclassification metrics

Table 1 Risk thresholds for common outcomes (outcomes examined in more than one of the 51 eligible studies)

Outcome	Risk thresholds (percent risk)	Risk duration	Studies (*n*)	Referenced thresholds (*n*)	Appropriate threshold use[a] (*n*)
Coronary heart disease	<5, 5–10, 10–20, >20	10 year	5	2	0
	<10, 10–20, >20	10 year	2	1	1
	<6, 6–20, >20	10 year	1	1	0
	<10, 10–25, 25–30, >30	10 year	1	1	0
	<5, 5–10, 10–15, >15	10 year	1	1	1
Composite CVD	<5, 5–10, 10–20, >20	10 year	4	2	0
	<10, 10–20, >20	10 year	3	2	0
	<6, 6–20, >20	10 year	1	1	0
	<20, >20	10 year	1	1	1
	<4, 4–7.5, >7.5	5 year	1	0	0
	<12, 12–40, >40	30 year	1	1	1
CVD mortality	<10, 10–20, >20	10 year	1	1	0
	<5, >5	10 year	1	0	0
	<5, 5–10, >10	10 year	1	1	1
	<6, 6–20, >20	10 year	1	0	0
	>1, 1–3, 3–6, >6	5 year	1	0	0
	<10, 10–20, >20	5 year	1	1	0
Type 2 diabetes	<2, 2–8, >8	10 year	1	0	0
	<5, 5–10, 10–15, >15	10 year	1	0	0
	<10, 10–20, >20	23.5* year	1	0	0
	Below/above average of baseline model	Not clear	1	0	0
All-cause mortality	<10, 10–20, >20	10 year	1	1	0
	<5, 5–10, 10–20, >20	10 year	1	0	0
	<10, 10–25, 25–35, >35	5 year	1	0	0
	<2.5, 2.5–5, 5–7, >7 (men); <1.3, 1.3–2.5, 2.5–3.8, >3.8 (women)	10 year	1	0	0
Mortality or MI in patients with previous CVD	<4.4, 4.4–13.0, >13.0	25 month	1	0	0
	<15, 15–72, >72	18 month	1	0	0
	<10, 10–20, >20	5 year	1	0	0
Fracture	>5, 5–15, >15	10 year	1	0	0
	>10, 10–20, >20	10 year	1	0	0
Mortality in end-stage liver disease patients	<10, 10–19, 20–30, 30–40, >40	90 day	1	0	0
	<5, 5–10, >10	1 year	1	0	0
Mortality in heart failure patients	<10, 10–30, >30	1 year	1	0	0
	<15, >15	5 year	1	1	0

[a]Thresholds used are those proposed by reference or authors have provided justification for use of alternative thresholds. MI: myocardial infarction. *median follow-up.

included liver transplantation, heart transplantation, bariatric surgery and thrombolysis/palliative care; two studies did not clarify what specific intervention they implied.

Model calibration and reclassification analysis

Use of calibration and reclassification among the 51 studies is presented in Figure 1. Overall 27 studies (53%) claimed that they had examined the calibration

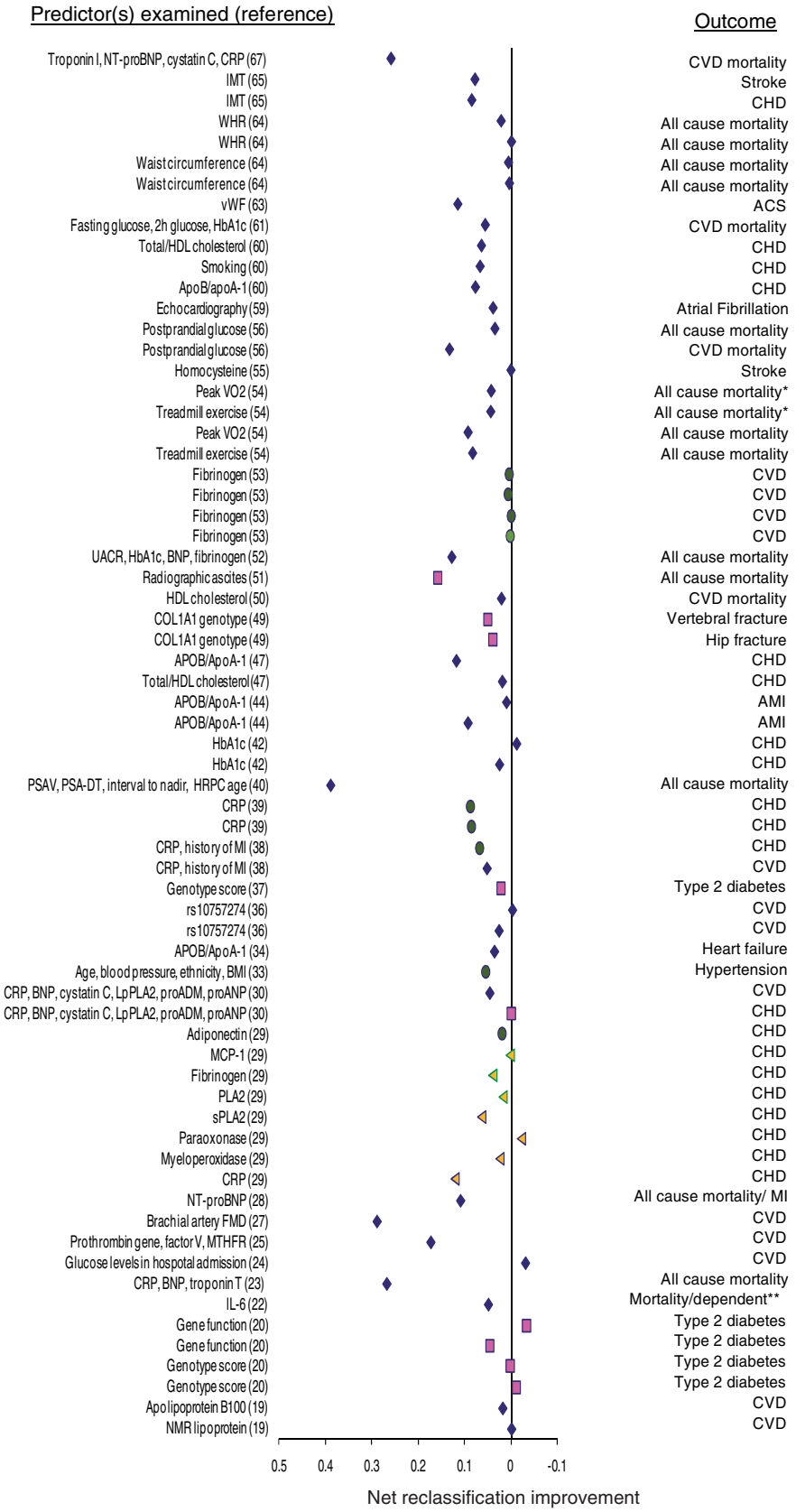

Figure 2 Extracted net reclassification index estimates from 67 analyses. Different markers indicate whether studies used baseline models that had been previously described, used as described, and had the same outcomes as originally intended

(continued)

of the baseline and baseline plus additional predictor models.

The NRI was calculated in 38 studies (70 analyses) among 48 studies published after 2008 when the NRI methodology was described (Figure 1). Three of these studies (three analyses) did not mention the actual NRI but only commented that the result was statistically significant/ non-significant. In 27 studies, the NRI was calculated based on time-to-event data (Cox regression); three of them reported complete (no loss of) follow-up.[75] The NRI ranged from −3.2 to 39% (Figure 2). With seven exceptions, all analyses suggested positive NRI. Twenty-six of the 58 NRIs with information on significance level were nominally statistically significant (all >0).

From 48 studies published after 2008, 19 studies (34 analyses) calculated the IDI and two studies (10 analyses) the relative IDI (Figure 1). The IDI ranged from 0 to 7.8% (Figure 3). Twenty-nine values were nominally statistically significant.

Reclassification estimates according to design features

For 22 NRI estimates and for four IDI estimates, baseline models had been previously described, used as described and had the same outcomes as originally intended (adequate baseline model). For 16 NRI estimates and for 1 IDI estimate the risk thresholds used were the same as the originally described references or they had justified the use of alternative thresholds (adequate threshold use) (Figures 2 and 3).

For 31 studies (58 analyses), the CIs of the NRI estimates were either provided by the study or could be calculated from the data provided in the paper. From those, the summary NRI was 0.030 (95% CI 0.014–0.045) for 9 studies (19 analyses) with adequate baseline models vs 0.016 (95% CI 0.011–0.022) in other studies ($P = 0.11$) (Supplementary Figure 2). The summary NRI was 0.009 (95% CI 0.003–0.014) in five studies (16 analyses) with adequate thresholds vs 0.030 (95% CI 0.021–0.039) in other studies ($P < 0.0001$) (Supplementary Figure 3). Only six and four NRI estimates of studies with adequate baseline models or adequate use of thresholds, respectively, were statistically significant. Qualitatively similar results were obtained when NRI was summarized only among studies which examined CHD or CVD as their outcome. From those, the summary NRI was 0.02 (95% CI 0.005–0.035) for 5 studies (16 analyses) with adequate baseline models vs 0.007 (95% CI 0.002–0.013) in other studies ($P = 0.11$). The summary NRI was 0.007 (95% CI 0.001–0.012) in 4 studies (15 analyses) with adequate thresholds vs 0.019 (95% CI 0.005–0.034) in other studies ($P = 0.10$). Too few IDI estimates had adequate use of baseline models or adequate use of thresholds to allow similar meaningful comparisons.

Overall, most [$n = 38$ (75%)] studies claimed that their results supported improved classification of the additional predictor and this did not differ for studies with or without adequate baseline models (10/14 vs 28/37) or studies with or without adequate risk thresholds (5/7 vs 33/44).

Discussion

In this empirical evaluation, the majority of studies claimed improved reclassification of a candidate predictor over and above established risk factors. However, most studies used baseline models which were not previously described, or were used differently or had different outcomes from those originally intended. Most studies used risk thresholds for reclassification that were not referenced or used as defined in the reference; and most studies used thresholds that were not linked to any management decisions. Moreover, almost half of the studies did not report on the calibration of the examined models and did not provide information on correct/incorrect reclassification percentages. Lack of adequate use of risk thresholds was associated with larger estimates of improved reclassification. Studies with adequate use of risk thresholds documented very limited reclassification ability.

Improved reclassification means making progress beyond what can be achieved with information already available from well-established traditional risk factors. However, most studies assessed here chose to show improved reclassification over models that were built for the first time or models for which it was unclear if they had ever been described before. Improved reclassification of a risk factor over and above such models may be appropriate as an exploratory analytical exercise, but it has questionable clinical value. Even among previously described models, some were well validated and the most widely used in the literature for the specific outcome (e.g. FRS for CHD), whereas others were less well validated or have

Figure 2 Continued
(pink/square markers), whether risk thresholds used were the same as the originally described references or authors had justified the use of alternative thresholds (green/circle markers) or both (orange/triangle markers). IMT: Intima media thickness; WHR: waist-hip ratio; vWF: von Willebrand factor; HDL: high density lipoprotein; UACR: urine albumin-creatinine ratio; BNP: B-type natriuretic peptide; APOB/APO-A1: apolipoprotein-B to apolipoprotein-AI ratio; PSAV: prostate-specific antigen velocity; PSA-DT: prostate-specific antigen doubling time; HRPC: hormone-refractory prostate cancer; CRP: C-reactive protein; MI: myocardial infarction; MCP-1: Monocyte chemotactic protein-1. *All cause mortality or United Network for Organ Sharing status 1 hearttransplantation. **Dependent on others after stroke

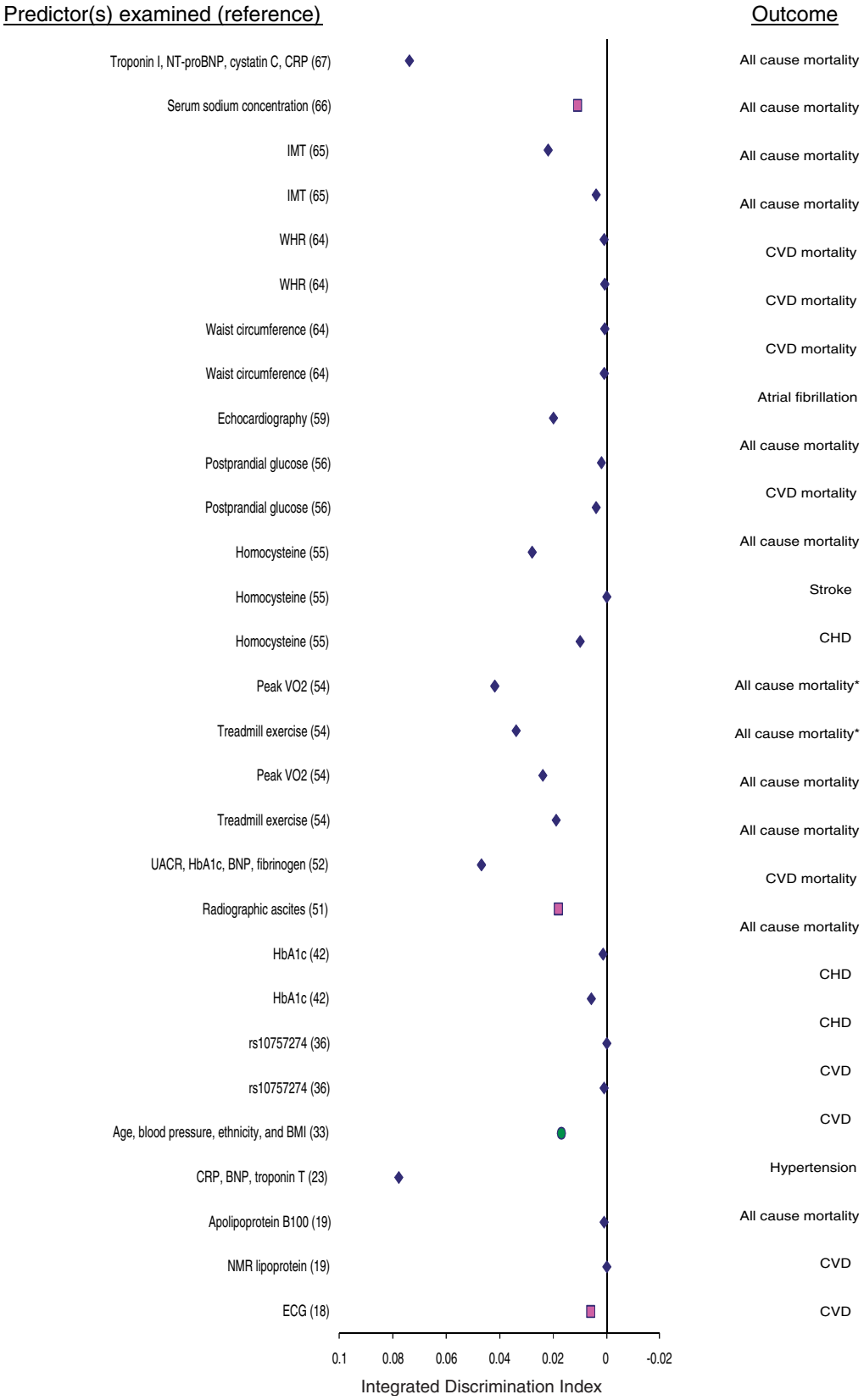

Figure 3 Extracted integrated discrimination improvement estimates from 29 analyses. Different markers indicate whether risk thresholds used were the same as the originally described references or authors have justified the use of alternative thresholds (adequate threshold: square markers) or whether studies also used baseline models that had been previously described, used as described, and had the same outcomes as originally intended (adequate baseline model: circle markers). Abbreviations are similar to Figure 2. *All cause mortality or United Network for Organ Sharing status 1 hearttransplantation

several contesters that may have better performance for predicting the same outcome. For example, other models for fracture risk[5] have been better validated than the nomogram used in Tran *et al*.[49] For several outcomes examined, e.g. type 2 diabetes, several predictive models have been described[76] but have not been externally validated or linked to treatment decisions.

Reclassification is sensitive to the choice of risk thresholds and reclassification capacity has been shown to increase dramatically with higher numbers of risk categories used.[77] In a recent simulation study, for the same discrimination accuracy (AUC), the total reclassification ranged from 0 to 22.5% depending on the selected cutoff thresholds.[78] Here, most examined studies used reclassification as another test of predictive ability without the thresholds having any relevance for medical decisions. Thresholds were chosen based on clinically irrelevant aspects, e.g. risk distribution;[32,40] sometimes the actual thresholds were not even mentioned.[63] Examined studies have shown inappropriate use of risk thresholds even for prediction of CHD, an outcome linked to well-established clinically relevant risk thresholds.[2] We documented empirically that inappropriate use of risk thresholds was associated with larger estimates of reclassification. We suspect that some of the large estimates of NRI may be spuriously inflated. There are many possibilities as to how this could happen, ranging from the expected inflation of classification metrics when a not-previously-validated model is fitted to the data, and spurious inflation when many analyses are performed using different thresholds, definitions or models, or combinations are used and only the best results are reported. Studies which used risk thresholds that were referenced and used as referenced or justified the use of alternative thresholds had an average NRI of 0.7%. Whereas there is no consensus on what is a large enough NRI, a value of 0.7% suggests that the relative proportions of patients whose prediction shifts in the right direction minus the proportions of those who shift in the wrong direction is only 0.7%. This is probably a tiny improvement, even if the available treatments are very effective. Data were too limited to investigate the relationship between methodological shortcomings and IDI estimates, but the clinical interpretation of IDI values is even more difficult and the values that we recorded were also generally small, with few exceptions.

Previous evaluations have shown that calibration is rarely examined in studies examining risk prediction.[10] Reclassification estimates (NRI and IDI) depend on model calibration.[11] In our sample of studies, half of examined studies did not report on calibration of the examined models. Lack of fit was rare among those that examined calibration, but it is not possible to exclude selective reporting bias leading to lack of reporting of poor calibration results.

Our study has some limitations. In particular, we used a sampling strategy that was systematic but was also driven by convenience. We have captured only a fraction of the studies that have done some reclassification exercise to date. Some articles might have preformed eligible analyses without mentioning reclassification results in the abstract and without citing the Pencina *et al*. paper.[11] However, it is unlikely that such studies would be methodologically more rigorous. If anything, studies with in passim mention of reclassification or lack of citation of standard methods might suffer even more from methodological drawbacks. In fact, the studies that we analysed may be a more sophisticated and higher quality sample of investigations in predictive medicine. Studies of prognostic factors have repeatedly been shown to have major deficiencies in methods and reporting.[10,14,79–83] A large component of our analysis was based on NRI and IDI measures as these are the most frequently used reclassification metrics. Other methods to assess reclassification such as the reclassification calibration statistic or the risk distribution curves had infrequent use in the studies that we examined and we would not have been able to describe their use based on limited numbers.[12,77,84–86]

Overall, we suggest that systematic efforts should be undertaken to put the predictive literature into some order[81,82] and identify for each disease and outcome what is the current best evidence for the best predictive models, risk thresholds and different treatment choices that are dictated by these thresholds. Such an effort would be a major undertaking, given the vastness of the data and the current lack of standardization in much of the corpus of predictive research. In scope, this effort may be compared with the task of the Cochrane Collaboration on systematic reviews for health care.[87] In particular, risk reclassification is an important tool in the assessment of the clinical relevance of a risk factor. Appropriate methodology and analysis are vital to avoid spurious claims of improved prediction and results of limited interpretability and misleading clinical inferences.

Supplementary data

Supplementary data are available at *IJE* online.

Conflict of interest: None declared.

KEY MESSAGES

- The majority of studies claimed improved reclassification of a candidate predictor over and above established risk factors.
- Most studies used baseline models which were not previously described, or were used differently or had different outcomes from those originally intended.
- Most studies used risk thresholds for reclassification that were not referenced or used as defined in the reference; and most studies used thresholds that were not linked to any management decisions.
- Reclassification studies would benefit from more rigorous methodological standards; otherwise claims for improved reclassification may remain spurious.

References

[1] Ioannidis JP. Limits to forecasting in personalized medicine: an overview. *Int J Forecast* 2009;**25:**773–83.

[2] Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP). Expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *JAMA* 2001; **285:**2486–97.

[3] Knaus WA, Wagner DP, Draper EA *et al*. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991;**100:**1619–36.

[4] Thompson IM, Ankerst DP, Chi C *et al*. Assessing prostate cancer risk: results from the prostate cancer prevention trial. *J Natl Cancer Inst* 2006;**98:**529–34.

[5] Kanis JA. Diagnosis of osteoporosis and assessment of fracture risk. *Lancet* 2002;**359:**1929–36.

[6] Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/ North American multicenter study. *JAMA* 1993;**270:** 2957–63.

[7] Gail MH, Brinton LA, Byar DP *et al*. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;**81:**1879–86.

[8] Kamath PS, Wiesner RH, Malinchoc M *et al*. A model to predict survival in patients with end-stage liver disease. *Hepatology* 2001;**33:**464–70.

[9] Allison M. Is personalized medicine finally arriving? *Nat Biotech* 2008;**26:**509–17.

[10] Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;**302:**2345–52.

[11] Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27:**157–72.

[12] Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med* 2008;**149:**751–60.

[13] Pepe MS, Janes H, Gu JW. Letter by Pepe et al regarding article, ''Use and misuse of the receiver operating characteristic curve in risk prediction''. *Circulation* 2007;**116:**e132.

[14] Ioannidis JP, Tzoulaki I. What makes a good predictor?: the evidence applied to coronary artery calcium score. *JAMA* 2010;**303:**1646–47.

[15] Lau J, Ioannidis JPA, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med* 1997;**127:**820–26.

[16] Ioannidis JPA, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ* 2008;**336:**1413–15.

[17] Hamer M, Chida Y, Stamatakis E. Association of very highly elevated C-reactive protein concentration with cardiovascular events and all-cause mortality. *Clin Chem* 2010;**56:**132–35.

[18] Gorodeski EZ, Ishwaran H, Blackstone EH, Lauer MS. Quantitative electrocardiographic measures and long-term mortality in exercise test patients with clinically normal resting electrocardiograms. *Am Heart J* 2009;**158:** 61–70.

[19] Mora S, Otvos JD, Rifai N, Rosenson RS, Buring JE, Ridker PM. Lipoprotein particle profiles by nuclear magnetic resonance compared with standard lipids and apolipoproteins in predicting incident cardiovascular disease in women. *Circulation* 2009;**119:**931–39.

[20] Talmud PJ, Hingorani AD, Cooper JA *et al*. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ* 2010; **340:**b4838.

[21] Sacco RL, Khatri M, Rundek T *et al*. Improving global vascular risk prediction with behavioral and anthropometric factors. The multiethnic NOMAS (Northern Manhattan Cohort Study). *J Am Coll Cardiol* 2009;**54:** 2303–11.

[22] Whiteley W, Jackson C, Lewis S *et al*. Inflammatory markers and poor outcome after stroke: a prospective cohort study and systematic review of interleukin-6. *PLoS Med* 2009;**6:**e1000145.

[23] Dunlay SM, Gerber Y, Weston SA, Killian JM, Redfield MM, Roger VL. Prognostic value of biomarkers in heart failure: application of novel methods in the community. *Circ Heart Fail* 2009;**2:**393–400.

[24] Correia LC, Rocha MS, Bittencourt AP *et al*. Does acute hyperglycemia add prognostic value to the GRACE score in individuals with non-ST elevation acute coronary syndromes? *Clin Chim Acta* 2009;**410:**74–78.

[25] Pezzini A, Grassi M, Del ZE *et al*. Common genetic markers and prediction of recurrent events after ischemic stroke in young adults. *Neurology* 2009;**73:**717–23.

[26] Mihaescu R, van Hoek M, Sijbrands EJ *et al*. Evaluation of risk prediction updates from commercial genome-wide scans. *Genet Med* 2009;**11:**588–94.

[27] Yeboah J, Folsom AR, Burke GL *et al*. Predictive value of brachial flow-mediated dilation for incident cardiovascular events in a population-based study: the multi-ethnic study of atherosclerosis. *Circulation* 2009; **120:**502–9.

[28] Eggers KM, Lagerqvist B, Venge P, Wallentin L, Lindahl B. Prognostic value of biomarkers during and after non-ST-segment elevation acute coronary syndrome. *J Am Coll Cardiol* 2009;**54:**357–64.

[29] Rana JS, Cote M, Despres JP *et al*. Inflammatory biomarkers and the prediction of coronary events among people at intermediate risk: the EPIC-Norfolk prospective population study. *Heart* 2009;**95:**1682–87.

[30] Melander O, Newton-Cheh C, Almgren P *et al*. Novel and conventional biomarkers for prediction of incident cardiovascular events in the community. *JAMA* 2009;**302:**49–57.

[31] Lin HJ, Lee BC, Ho YL *et al*. Postprandial glucose improves the risk prediction of cardiovascular death beyond the metabolic syndrome in the nondiabetic population. *Diabetes Care* 2009;**32:**1721–26.

[32] Moayyeri A, Kaptoge S, Dalzell N *et al*. The effect of including quantitative heel ultrasound in models for estimation of 10-year absolute risk of fracture. *Bone* 2009; **45:**180–84.

[33] Paynter NP, Cook NR, Everett BM, Sesso HD, Buring JE, Ridker PM. Prediction of incident hypertension risk in women with currently normal blood pressure. *Am J Med* 2009;**122:**464–71.

[34] Holme I, Strandberg TE, Faergeman O *et al*. Congestive heart failure is associated with lipoprotein components in statin-treated patients with coronary heart disease insights from the incremental decrease in end points through aggressive lipid lowering trial (IDEAL). *Atherosclerosis* 2009;**205:**522–27.

[35] Meneveau N, Schiele F, Seronde MF *et al*. Anemia for risk assessment of patients with acute coronary syndromes. *Am J Cardiol* 2009;**103:**442–47.

[36] Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann Intern Med* 2009;**150:**65–72.

[37] Meigs JB, Shrader P, Sullivan LM *et al*. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* 2008;**359:**2208–19.

[38] Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds risk score for men. *Circulation* 2008;**118:**2243–51, 4p.

[39] Shah T, Casas JP, Cooper JA *et al*. Critical appraisal of CRP measurement for the prediction of coronary heart disease events: new data and systematic review of 31 prospective cohorts. *Int J Epidemiol* 2009;**38:**217–31.

[40] Robinson D, Sandblom G, Johansson R *et al*. PSA kinetics provide improved prediction of survival in metastatic hormone-refractory prostate cancer. *Urology* 2008;**72:** 903–7.

[41] Fowkes FG, Murray GD, Butcher I *et al*. Ankle brachial index combined with Framingham risk score to predict cardiovascular events and mortality: a meta-analysis. *JAMA* 2008;**300:**197–208.

[42] Simmons RK, Sharp S, Boekholdt SM *et al*. Evaluation of the Framingham risk score in the European Prospective Investigation of Cancer-Norfolk cohort: does adding glycated hemoglobin improve the prediction of coronary heart disease events? *Arch Intern Med* 2008;**168:**1209–16.

[43] Kathiresan S, Melander O, Anevski D *et al*. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med* 2008;**358:**1240–49.

[44] Talmud PJ, Cooper JA, Palmen J *et al*. Chromosome 9p21.3 coronary heart disease locus genotype and prospective risk of CHD in healthy middle-aged men. *Clin Chem* 2008;**54:**467–74.

[45] Holme I, Aastveit AH, Jungner I, Walldius G. Relationships between lipoprotein components and risk of myocardial infarction: age, gender and short versus longer follow-up periods in the Apolipoprotein MOrtality RISk study (AMORIS). *J Intern Med* 2008;**264:** 30–38.

[46] Hallan S, Astor B, Romundstad S, Aasarod K, Kvenild K, Coresh J. Association of kidney function and albuminuria with cardiovascular mortality in older vs younger individuals: The HUNT II Study. *Arch Intern Med* 2007;**167:** 2490–96.

[47] Ingelsson E, Schaefer EJ, Contois JH *et al*. Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. *JAMA* 2007;**298:** 776–85.

[48] Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med* 2006;**145:**21–29.

[49] Tran BN, Nguyen ND, Center JR, Eisman JA, Nguyen TV. Enhancement of absolute fracture risk prognosis with genetic marker: the collagen I alpha 1 gene. *Calcif Tissue Int* 2009;**85:**379–88.

[50] Cooney MT, Dudina A, De BD *et al*. How much does HDL cholesterol add to risk estimation? A report from the SCORE Investigators. *Eur J Cardiovasc Prev Rehabil* 2009; **16:**304–14.

[51] Somsouk M, Guy J, Biggins SW, Vittinghoff E, Kohn MA, Inadomi JM. Ascites improves upon [corrected] serum sodium plus [corrected] model for end-stage liver disease (MELD) for predicting mortality in patients with advanced liver disease. *Aliment Pharmacol Ther* 2009;**30:** 741–48.

[52] Kizer JR, Krauser DG, Rodeheffer RJ *et al*. Prognostic value of multiple biomarkers in American Indians free of clinically overt cardiovascular disease (from the Strong Heart Study). *Am J Cardiol* 2009;**104:**247–53.

[53] Woodward M, Tunstall-Pedoe H, Rumley A, Lowe GD. Does fibrinogen add to prediction of cardiovascular disease? Results from the Scottish Heart Health Extended Cohort Study. *Br J Haematol* 2009;**146:**442–46.

[54] Hsich E, Gorodeski EZ, Starling RC, Blackstone EH, Ishwaran H, Lauer MS. Importance of treadmill exercise time as an initial prognostic screening tool in patients with systolic left ventricular dysfunction. *Circulation* 2009;**119:**3189–97.

[55] Sun Y, Chien KL, Hsu HC, Su TC, Chen MF, Lee YT. Use of serum homocysteine to predict stroke, coronary heart

disease and death in ethnic Chinese. 12-year prospective cohort study. *Circ J* 2009;**73:**1423–30.

[56] Cortigiani L, Bombardini T, Corbisiero A, Mazzoni A, Bovenzi F, Picano E. The additive prognostic value of end-systolic pressure-volume relation in patients with diabetes mellitus having negative dobutamine stress echocardiography by wall motion criteria. *Heart* 2009;**95:** 1429–35.

[57] Pencina MJ, D'Agostino RB Sr, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the Framingham heart study. *Circulation* 2009; **119:**3078–84.

[58] Cortigiani L, Sicari R, Bigi R, Landi P, Bovenzi F, Picano E. Impact of gender on risk stratification by stress echocardiography. *Am J Med* 2009;**122:**301–9.

[59] Schnabel RB, Sullivan LM, Levy D *et al*. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet* 2009; **373:**739–45.

[60] Holme I, Cater NB, Faergeman O *et al*. Lipoprotein predictors of cardiovascular events in statin-treated patients with coronary heart disease. Insights from the Incremental Decrease in End-points through Aggressive Lipid-lowering trial (IDEAL). *Ann Med* 2008;**40:**456–64.

[61] Barr EL, Boyko EJ, Zimmet PZ, Wolfe R, Tonkin AM, Shaw JE. Continuous relationships between non-diabetic hyperglycaemia and both cardiovascular disease and all-cause mortality: the Australian Diabetes, Obesity, and Lifestyle (AusDiab) study. *Diabetologia* 2009;**52:**415–24.

[62] Lyssenko V, Jonsson A, Almgren P *et al*. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N Engl J Med* 2008;**359:**2220–32.

[63] Empana JP, Canoui-Poitrine F, Luc G *et al*. Contribution of novel biomarkers to incident stable angina and acute coronary syndrome: the PRIME Study. *Eur Heart J* 2008; **29:**1966–74.

[64] Pischon T, Boeing H, Hoffmann K *et al*. General and abdominal adiposity and risk of death in Europe. *N Engl J Med* 2008;**359:**2105–20.

[65] Chien KL, Su TC, Jeng JS *et al*. Carotid artery intima-media thickness, carotid plaque and coronary heart disease and stroke in Chinese. *PLoS One* 2008;**3:**e3435.

[66] Kim WR, Biggins SW, Kremers WK *et al*. Hyponatremia and mortality among patients on the liver-transplant waiting list. *N Engl J Med* 2008;**359:**1018–26.

[67] Zethelius B, Berglund L, Sundstrom J *et al*. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *N Engl J Med* 2008;**358:**2107–16.

[68] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;**97:** 1837–47.

[69] Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med* 2007;**167:**1068–74.

[70] Rahman M, Simmons RK, Harding AH, Wareham NJ, Griffin SJ. A simple risk score identifies individuals at high risk of developing type 2 diabetes: a prospective cohort study. *Fam Pract* 2008;**25:**191–96.

[71] Counsell C, Dennis M, McDowall M, Warlow C. Predicting outcome after acute and subacute stroke: development and validation of new prognostic models. *Stroke* 2002;**33:**1041–47.

[72] Chobanian AV, Bakris GL, Black HR *et al*. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. *JAMA* 2003;**289:**2560–72.

[73] Conroy RM, Pyorala K, Fitzgerald AP *et al*. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;**24:**987–1003.

[74] Nguyen ND, Frost SA, Center JR, Eisman JA, Nguyen TV. Development of a nomogram for individualizing hip fracture risk in men and women. *Osteoporos Int* 2007;**18:** 1109–17.

[75] Steyerberg EW, Pencina MJ. Reclassification calculations for persons with incomplete follow-up. *Ann Intern Med* 2010;**152:**195–96.

[76] Mann DM, Bertoni AG, Shimbo D *et al*. Comparative validity of 3 diabetes mellitus risk prediction scoring models in a multiethnic US cohort: the multi-ethnic study of atherosclerosis. *Am J Epidemiol* 2010;**171:**980–98.

[77] Pepe MS, Gu JW, Morris DE. The potential of genes and other markers to inform about risk. *Cancer Epidemiol Biomarkers Prev* 2010;**19:**655–65.

[78] Mihaescu R, van Zitteren M, van Hoek M *et al*. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol* 2010;**172:** 353–61.

[79] Kyzas PA, axa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer* 2007;**43:**2559–79.

[80] Kyzas PA, axa-Kyza D, Ioannidis JP. Quality of reporting of cancer prognostic marker studies: association with reported prognostic effect. *J Natl Cancer Inst* 2007;**99:** 236–43.

[81] Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;**338:**b606.

[82] Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;**338:**b375.

[83] Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;**338:**b604.

[84] Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009;**150:** 795–802.

[85] Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: Extension to survival analysis. *Statist Med* 2010;**30:**22–38.

[86] Janes H, Pepe MS, Gu W. Are risk stratification tables the best way to evaluate model performance? *Ann Intern Med* 2009;**150:**428.

[87] The Cohrane Collaboration. http://www.cochrane.org/ 2010 (October 2010, date last accessed).