

Assessment of Claims of Improved Prediction Beyond the Framingham Risk Score

Ioanna Tzoulaki, PhD

George Liberopoulos, MD

John P. A. Ioannidis, MD

EMPHASIS ON IMPROVEMENTS IN prediction has become a hallmark of the quest for personalized and individualized medicine,¹ and the predictive literature is drawing increasing attention in medicine. However, the utility of this literature can be hampered by methodological limitations affecting design, analysis, and reporting.²⁻⁶ One challenge is to demonstrate that new candidate predictors can offer independent, incremental information beyond what is already known based on traditional risk factors. A sophisticated new predictor may have good predictive ability on its own but may not improve predictive ability further when simple, easy-to-measure traditional factors are already taken into account.

We assessed empirically a systematic sample of studies that evaluated various candidate prognostic factors in their ability to improve prediction of coronary heart disease (CHD) or other outcomes beyond what the Framingham risk score (FRS) can achieve.⁷ The FRS is one of the most thoroughly validated and widely used predictive scores in medical literature. Its original publication in 1998 is one of the most-cited articles across all biomedicine.⁷ The FRS was developed with robust methods and is intended to offer prospective risk assessment for CHD risk in men and women who do not have overt CHD.

See also Patient Page.

For editorial comment see p 2369.

Context With heightened interest in predictive medicine, many studies try to document information that can improve prediction of major clinical outcomes.

Objective To evaluate the reported design and analysis of studies that examined whether additional predictors improve predictive performance when added to the Framingham risk score (FRS), one of the most widely validated and cited clinical prediction scores.

Study Selection Two independent investigators searched 1908 articles citing the article that described the FRS in 1998 until September 2009 through the ISI Web of Knowledge database. Articles were eligible if they included any analyses comparing the predictive performance of the FRS vs the FRS plus some additional predictor for a prospectively assessed outcome.

Data Analyses We recorded information on FRS calculation, modeling of additional predictors, outcomes assessed, population evaluated, subgroup analysis documentation, and flaws in the methods that may have affected the reported improvements in predictive ability. We also evaluated the correlation of reported design and analysis features with the predictive model discrimination and improvements with the additional predictors.

Results We evaluated 79 eligible articles. Forty-nine studies (62%) did not calculate the FRS as it has been proposed, 15 (19%) modeled the additional predictor in more than 1 way and presented only the best-fit or area-under-the-curve (AUC) results for only 1 model, 41 (52%) did not examine the original outcome that the FRS was developed for, 33 (42%) studied a population different from what the FRS was intended for, and 25 (32%) claimed improved prediction in 1 subgroup but only 7 (9%) formally tested subgroup differences. Evaluation of independence in multivariable regressions, discrimination in AUC, calibration, and reclassification were reported in 77, 36, 7, and 7 studies, respectively, but these methods were adequately documented in only 60, 13, 4, and 2 studies, respectively. Overall, 63 studies (80%) claimed some improved prediction. Increase in AUC was larger when the predictive performance of the FRS was lower ($\rho = -0.57, P < .001$). Increase in AUC was significantly larger when evaluation of independence in multivariable regression or discrimination in AUC analysis was not adequately documented and when the additional predictor had been modeled in more than 1 way and only 1 model was reported for AUC.

Conclusion The majority of examined studies claimed that they found factors that could offer additional predictive value beyond what the FRS could achieve; however, most had flaws in their design, analyses, and reporting that cast some doubt on the reliability of the claims for improved prediction.

JAMA. 2009;302(21):2345-2352

www.jama.com

Risk is calculated based on age, blood pressure, total or low-density lipoprotein cholesterol level, high-density lipoprotein cholesterol level, smoking status, and presence of diabetes mellitus. We aimed to assess, among studies that tried to find additional risk factors that could improve prediction beyond the FRS,

whether the reported design and analysis were adequate or claims for improvement were susceptible to bias.

Author Affiliations are listed at the end of this article.
Corresponding Author: John P. A. Ioannidis, MD, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece (jioannid@cc.uoi.gr).

METHODS

Selection of Studies

We assembled studies that examined whether 1 or more factors other than those included in the FRS could improve the predictive performance of the FRS. Identifying all studies that have done such analyses is impractical, if not impossible. Instead, we assembled a systematic sample with an objective and readily replicable search strategy. Thus, we searched all articles that cited the 1998 *Circulation* article describing the FRS,⁷ herein called the “reference FRS article.” We used the ISI Web of Knowledge database, and the last search was updated in September 2009, yielding 1908 citations. We perused the title and abstract of each citing article. Two evaluators (I.T., G.L.) independently searched all articles with a third evaluator (J.P.A.I.) arbitrating their discrepancies. Potentially eligible articles were then scrutinized in full text.

Articles were eligible if they included 1 or more analyses comparing the predictive performance of the FRS against the predictive performance of the FRS plus some additional predictor for a prospectively assessed outcome. Eligible articles cited the reference FRS article either in the methods in describing what they used as baseline for the comparison or in some other section alluding to traditional risk factors or the FRS in a way that suggested that FRS variables were used as the baseline for the comparison (as long as the methods did not suggest that some other score or model was used). We included articles regardless of the types of additional predictors considered and how they were measured and handled in the analysis and regardless of whether the examined outcome was CHD as defined by the reference FRS article, CHD with different definitions, or some other outcome.

Moreover, we considered articles regardless of what analyses were reported to address incremental predictive ability (evaluation of independence in multivariable regressions with and without the additional predictors, evaluation of discrimination by calculation and comparison of area under the curve [AUC]

with and without the additional predictors in receiver operating characteristic [ROC] curves, evaluation of model calibration [accuracy of absolute risk prediction in various risk categories], evaluation of reclassification of participants into risk categories, or any other method addressing incremental predictive ability). We included studies regardless of whether they gave quantitative results or only qualitative statements comparing predictive models. Studies were included regardless of whether the eligible comparative analyses were the sole analyses or among several assessed in the same article.

We did not consider articles without data on actual patients (eg, decision analyses with simulated data); reviews; studies without prospective follow-up (eg, case-control and cross-sectional studies); studies that examined only the correlation/regression of additional predictors against the FRS or components thereof; studies that sought only to recalibrate the FRS in a different population, without considering additional predictors; and studies that compared the FRS with other risk scores without evaluating the additive value of these scores over and above the FRS.

When an article considered separately 2 or more predictors for their incremental predictive ability or more than 1 outcome, each was considered a different analysis. When several additional predictors were considered together, we did not separate them.

Definitions and Data Extraction

Two evaluators (I.T., G.L.) independently extracted data from eligible studies. Any inconsistencies were identified and cross-checked by both investigators with a third evaluator (J.P.A.I.) arbitrating any remaining discrepancies. For each eligible article, we recorded first author, journal, publication year, and sample size; information on the FRS, additional predictors, outcomes assessed, population evaluated, and methods of evaluation of predictive ability that may have affected study results; and inferences on whether the additional predictors improved prediction beyond the FRS.

The estimated baseline performance of the FRS alone and the estimated improvement by a candidate additional predictor may be affected in a number of ways: if the FRS is calculated or used suboptimally (not as originally developed and proposed), if the candidate additional predictor is modeled in a way that its performance is inappropriately exaggerated, if the examined outcome or population is not what the FRS was intended for, or if spurious claims are made for improved prediction only in specific subgroups of patients. We tried to capture these diverse possibilities.

We recorded the following aspects of suboptimal FRS calculation and use: whether any of the standard risk factors was deleted or modified or other factors added; whether it was not calculated based on the regression coefficients or points proposed for each factor in the reference FRS article (or the well-accepted minor variant version of the National Cholesterol Education Program Adult Treatment Panel III guidelines),⁸ but based on some other approach (eg, new regression coefficients developed from the data in the specific study); whether patients with specific FRS values were excluded, narrowing or altering the dynamic range; and whether the FRS was not used as a continuous score but in grouped categories.

For each additional candidate predictor, we noted whether it had been modeled in more than 1 way (eg, continuous and categorical, or different splits used in categorical analyses) and results for incremental predictive ability were presented for only 1 model. If so, we recorded whether the presented results were clearly the best fit (those with maximum estimated improvement in predictive ability) or this was unclear.

For outcomes, we recorded whether they were CHD or some completely different outcome. For CHD, we further recorded whether it was defined as in the reference FRS article (angina pectoris, myocardial infarction, coronary insufficiency, or CHD death) and whether outcome was assessed within 10 years of follow-up. For the population studied, we recorded whether it included partici-

pants of nonwhite ethnicity (>10%) or participants with overt history of CHD at start of follow-up. For subgroup analyses, we recorded whether claims were made for improved predictive performance only in a specific subgroup of patients without properly documenting that the improvement is different beyond chance ($P < .05$) in these subgroups vs other participants.

Assessing Incremental Predictive Ability

Several methods for the evaluation of prediction models have been proposed. For each article and additional predictor of interest, we noted which methods of evaluation of predictive ability were reported. For 4 methods that have been widely discussed and proposed in the CVD literature and beyond, we noted whether they had been adequately documented.

For multivariable regressions examining independence of the additional predictor, we considered documentation to be adequate when a test was used (eg, Bayes or Akaike information criteria) that penalized for including an additional variable in the model (fully adequate documentation) or when the investigators reported whether the additional predictor was independently statistically significant at $P < .05$ without addressing model parsimony (partially adequate documentation).

For the evaluation of discrimination in ROC curves,⁹ AUC (or C-index) analyses were considered adequately documented if there was description or reference to the method, provision of both AUC estimates (with and without the additional predictor), and provision of either confidence intervals (CIs) or a P value for the comparison of AUCs. Values for the AUC, 95% CIs, and P values were recorded when 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.

Calibration (accuracy) of the predictive model was considered adequately documented when model calibration was given both with and without the

additional predictor with a goodness-of-fit test.

Adequate documentation of reclassification required the following: use of standard categories of risk (<10%, 10%-20%, >20%)⁸ or justified use of different categories; reporting of how many participants changed risk categories for each type of change; and reporting of whether participants who had changed risk category had moved in the correct direction depending on whether they had an event (improved reclassification if moving to a higher predicted risk category for those who had an event and if moving to a lower predicted risk category for those who did not have an event).¹⁰

Finally, we estimated what proportion of studies claimed in interpreting

their findings that they had found an improvement in prediction.

Correlates of AUC Estimates and Improvement in AUC

We evaluated whether improvement in AUC (the difference in AUCs with vs without the additional predictor, Δ AUC) was larger when the reported AUC of FRS alone was lower. We calculated the Spearman correlation coefficient between Δ AUC and the AUC of FRS alone; in sensitivity analyses, we evaluated the correlation between Δ AUC and the mean of the AUC of FRS alone and the AUC of the FRS plus the additional predictors.

We also evaluated whether reported design and analysis features correlated with results on predictive per-

Table 1. Additions, Deletions, and Modifications of the Risk Factors Used for the Framingham Risk Score in 79 Eligible Articles Compared With the Framingham Risk Score Definition in Wilson et al⁷

| | No. of Studies |
|--|----------------|
| Additions | |
| Antihypertensive or lipid-lowering medication ^{22,26,29,55,58,64,67,74,79} | 9 |
| Obesity, body mass index, waist-hip ratio ^{22,24,25,33,35,37} | 6 |
| Race/ethnic group ^{20,26,31,36,40} | 5 |
| Triglycerides ^{27,30} | 2 |
| Alcohol intake ^{23,24} | 2 |
| Prevalent cardiovascular disease ^{34,38} | 2 |
| Social class ^{36,40} | 2 |
| Other risk factors ^{23,36} | 2 |
| Deletions | |
| Diastolic blood pressure ^{22,26,29,32,39,55,58,64,67,74,79, a} | 11 |
| Diabetes ^{27,30,32,35,90} | 5 |
| HDL cholesterol ^{27,30,35,36,40} | 5 |
| All ^{35, b} | 1 |
| Modifications | |
| Blood pressure definition | |
| Systolic blood pressure and antihypertensive definition ^{22,26,29,55,58,64,67,74,79, a} | 9 |
| History of hypertension, self-reported hypertension ^{17,19,21,24,90} | 5 |
| Other ^{24,32-34,39,65} | 6 |
| Definition of diabetes | |
| Fasting glucose >126 mg/dL ^{32,41,45,49,54,61,72,84, a} | 8 |
| History of diabetes, self-reported diabetes ^{28,82} | 2 |
| Lipid levels | |
| Use of cholesterol/HDL cholesterol ratio ^{22,29,32,39} | 4 |
| History of hyperlipidemia, self-reported hyperlipidemia ^{17,19,21,90} | 4 |
| Smoking | |
| Pack-years of smoking ³⁹ | 1 |
| Use of exsmokers category ^{28,72} | 2 |

Abbreviation: HDL, high-density lipoprotein.

^aThese definitions were not considered inappropriate when authors reported using the National Cholesterol Education Program Adult Treatment Panel III guidelines.

^bThis study used stepwise regression with backward selection; in the final model, all Framingham risk score variables were eliminated.

formance. Specifically, we evaluated whether the AUC with and without the additional predictors as well as the Δ AUC were different depending on whether the analysis involved adequate vs suboptimal calculation or use of the FRS; appropriate vs potentially exaggerated modeling of the additional predictor; CHD vs other outcome; appropriate population (white ethnicity, no baseline CHD) vs other target population; the whole population vs a subgroup; adequate vs inadequate documentation of independence in multivariable regression; and adequate vs inadequate documentation of discrimination in AUC analyses, as defined previously. Too few studies had performed calibration and reclassification evaluations to allow meaningful similar analyses. We used the Mann-Whitney U test to compare the point estimates in AUC and in Δ AUC between subgroups defined by these study features. A 2-sided *P* value $<.05$ was used to denote statistical significance. Statistical tests were performed with SPSS version 14 (SPSS Inc, Chicago, Illinois).

RESULTS

After exclusions (eFigure; see <http://www.jama.com>), 79 articles were eligible.¹¹⁻⁸⁹ The κ coefficient for eligibility between the 2 independent investigators on initial screening was 0.86. Main study characteristics are shown in eTables 1 and 2 (see <http://www.jama.com>).

Calculation and Use of the FRS

Twenty-four studies¹¹⁻³⁴ used different risk factors from those included in the reference FRS article. Additions, deletions, and modifications of risk factors appear in TABLE 1. Twenty-three studies^{11,13,17-23,25-38} did not calculate the FRS but simply used variables that they claimed corresponded to FRS factors. Three articles³⁹⁻⁴¹ narrowed the dynamic range by excluding participants from further analysis based on FRS values. In 19³⁹⁻⁵⁷ of 56 studies that calculated a continuous score, the score was categorized and the continu-

ous information not used. Overall, 49 studies^{11-37,39-56,58-61} (62%) calculated or used the FRS suboptimally for at least 1 aspect described here.

Additional Predictors

In total, 86 different predictors were assessed in 192 analyses; between 1 and 51 different predictors were assessed per article (eTable 1). Fifteen studies (19%) examined different models of analysis; 4 studies^{11,40,61,62} reported the model with the best fit, and 11 studies^{24,25,31,53,54,57,63-67} reported only the results of the AUC analysis for one model without clarifying whether it was the best-fitting one.

Outcomes

Forty-one studies§ (52%) did not examine CHD as an outcome of interest in any analysis. Various other outcomes were assessed (eTable 3).

Follow-up ranged from 2 to 32 years (median, 7.1 years); only 23 studies (29%) had a mean or median follow-up time of 10 years or longer and 35 studies|| (44%) a mean or median follow-up time of 8 years or longer.

Study Population

A population of exclusively white ethnicity was examined in 39 studies (49%).¶ Among the remaining studies, 24 studies# included more than 10% nonwhite participants (2 had exclusively nonwhite participants), and for 16 studies, population ethnicity was unclear. In 18 studies (23%),** participants with overt CHD at baseline were included. Overall, 33 studies (42%) clearly included more than 10% nonwhite participants or patients with CHD at baseline.

§References 11, 12, 15-17, 19, 27-37, 39, 43, 44, 46, 48, 52, 55-57, 61, 62, 77-89.

||References 14, 16-19, 21, 23, 24, 26-28, 30, 35, 38, 41, 43, 45, 46, 49, 53, 54, 58, 61, 65, 69, 71, 75-83.

¶References 12, 13, 16, 17, 21, 23, 24, 26, 30-33, 37, 39, 40, 47-53, 57, 61, 63-65, 67, 69, 71, 72, 74, 76, 78, 80, 83-86.

#References 14, 20, 25, 27-29, 34, 35, 38, 42, 45, 54, 56, 58, 60, 66, 68, 70, 73, 75, 77, 81, 82, 89.

**References 28-34, 36, 37, 41, 55-57, 63, 79, 83-85.

Subgroup Analyses

Thirty-nine studies (51%) stratified the analysis examining the additive predictive value of a novel risk factor in different subgroups. Subgroup analysis was defined according to FRS categories (15 studies††), categories of other factors (16 studies‡‡), or a combination (7 studies^{16,40,45,48,54,62,69}). Improved prediction in one subgroup vs others was claimed in 25§§ of those 39 studies (64%). Only 7 studies^{12,25,28,45,49,73,75} (18%) presented a formal test for differences between subgroups.

Assessing Incremental Predictive Ability

Of the 79 eligible studies, 77 reported some multivariable regression analysis for at least 1 additional predictor and at least 1 outcome. The other 2 studies^{52,56} performed only univariate regressions of the additional predictor stratified by FRS category. Also, 36 studies|||| used ROC analysis, 7 studies^{11,14,23,24,61,67,85} tested the calibration of multivariable regressions, and 7 studies^{24,54,57,61,64,67,76} described reclassification analysis. Documentation of these methods appears in TABLE 2.

Overall, 63 articles¶¶ (80%) claimed improved prediction for at least 1 predictor and 1 outcome.

AUC Analyses

Twenty-nine studies reported AUC estimates for both FRS alone and the FRS with additional predictors with data on 88 such pairs of data. The AUC of FRS alone and the FRS with additional predictors ranged from 0.50 to 0.83 (median, 0.74) and 0.57 to 0.84 (median, 0.75), respectively. There was a strong inverse correlation between the Δ AUC and the baseline AUC of FRS alone

††References 39, 42-44, 52, 53, 55-57, 59, 66, 71, 74, 86, 88.

‡‡References 12, 14, 15, 25-28, 46, 49, 58, 60, 63, 68, 75, 76, 87.

§§References 12, 14-16, 25, 26, 28, 40, 42, 44, 45, 48, 54, 55, 58-60, 66, 69, 71, 73-76, 87.

||||References 11, 12, 14-16, 20, 21, 24-26, 31, 33, 54, 57-61, 63-68, 70, 71, 73, 74, 76-78, 81, 82, 85, 88, 89.

¶¶References 11-19, 21, 22, 24, 27-30, 32-36, 38-45, 47-55, 57-59, 61-63, 66, 68-70, 72-76, 78-81, 83-88.

(Spearman correlation coefficient, -0.57 ; $P < .001$) and between ΔAUC and the mean AUC of the FRS alone and the FRS with additional predictors (correlation coefficient, -0.47 ; $P < .001$). Absolute improvements in AUC of at least 0.05 were seen only when FRS achieved an AUC of 0.72 or less (FIGURE).

TABLE 3 shows the median AUC values and the estimated improvement in prediction (ΔAUC) when the data were classified according to prespecified features of study design and analysis. The AUC estimates were significantly higher in studies with suboptimal FRS calculation or use vs those where the FRS was calculated and used as proposed; in studies that used CHD rather than other outcomes; and in those where the evaluation of independent information in multivariable regression was adequately documented vs those that did not adequately document this aspect. Also, the ΔAUC was significantly larger when multivariable regression or ROC analysis was not adequately documented and when the additional predictor had been modeled in more than 1 way but only 1 model was reported for AUC analyses.

COMMENT

In this empirical evaluation, the majority of examined studies claimed that they found factors that could offer additional predictive value beyond what the FRS could achieve. However, most studies had flaws in their design, analyses, and reporting that cast doubt on the reliability of the claims of improved prediction. Furthermore, some methodological limitations were associated with larger estimates of reported improvements in predictive ability; studies without such limitations showed on average no improvement in AUC with the additional candidate predictors.

Most studies examining the additive predictive value of a risk factor with the FRS did so by presenting statistically significant associations with the outcome after the FRS or its components were included in multivariable regression models. However, a P value alone usually offers weak support for

Table 2. Documentation of Methods of Assessing Incremental Predictive Ability

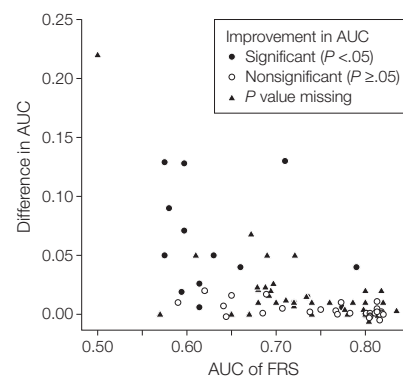
| | No. (%) |
|---|-----------|
| Adequate documentation of multivariable regressions (n = 77 studies) | |
| (a) Information on whether additional predictor is significant at $< .05$ level | 57 (74.0) |
| (b) Results of a test that penalizes for the inclusion of additional predictor | 15 (19.5) |
| Bayes or Akaike information criteria | 6 (7.8) |
| Likelihood ratio test | 9 (11.7) |
| Both a and b | 12 (15.6) |
| Appropriate documentation (a or b) | 60 (77.9) |
| Adequate documentation of AUC in ROC analysis (n = 36 studies) | |
| (a) Described method used to compare ROC curves | 20 (55.6) |
| (b) Presented the AUC values with and without the additional predictor | 29 (80.6) |
| (c) Presented CIs of AUC values with and without additional predictor | 11 (30.6) |
| (d) Presented P value for comparison | 18 (50.0) |
| Appropriate documentation (a and b and [c or d]) | 13 (36.1) |
| Adequate documentation of calibration (n = 7 studies) | |
| Presented results comparing models with and without additional predictor | 4 (57.1) |
| Adequate documentation of reclassification analysis (n = 7 studies) | |
| (a) Use of standard categories of risk | 2 (28.6) |
| (b) Justified use of other categories of risk | 2 (28.6) |
| (c) Reported the number of patients changing categories in correct direction | 5 (71.4) |
| Appropriate documentation ([a or b] and c) | 2 (28.6) |

Abbreviations: AUC, area under the curve; CIs, confidence intervals; ROC, receiver operating characteristic.

credibility and provides no information on model calibration and discrimination.^{91,92} Calibration, the important ability of the model to give predictions of absolute risk that are commensurate with levels seen in reality, was rarely reported. Discrimination, how well the model can separate those who do and do not experience an outcome, was reported in about a third of the studies via AUC analysis. However, even then, reporting of AUC analyses was often inadequate. Few studies assessed the ability of the additional risk factor to reclassify individuals into risk categories and only 2 provided adequate documentation. Reclassification analyses are essential for understanding whether a new model can alter management decisions.^{10,93} The sparse use of these methods, partly explained by their recent introduction, and suboptimal reporting diminish the clinical utility of this literature.

Considerable absolute improvements in AUC were reported only in studies where the FRS performed poorly, below what is typically expected of this widely used and carefully validated score. This may be due to chance (regression to the mean),

Figure. Correlation Between Predictive Ability of the Framingham Risk Score and Change in Predictive Ability After Inclusion of Additional Predictors



Association between area under the receiver operating curve (AUC) of Framingham risk score (FRS) and difference in AUC between AUC of FRS alone and AUC of FRS plus inclusion of additional predictor in 88 analyses. Different markers indicate whether the improvement in AUC was found to be nominally statistically significant ($P < .05$) (17 analyses) or not statistically significant (30 analyses) or that information on level of significance was missing in the article (41 analyses).

genuine differences in study characteristics, or a conglomerate of diverse biases that decrease the predictive performance of the FRS. The baseline predictive performance of the FRS was

Table 3. Median AUC Values and Δ AUC According to Different Aspects of Study Design and Analysis^a

| | No. | AUC FRS, Median (IQR) | P Value ^b | AUC FRS+AD, Median (IQR) | P Value ^b | Difference in AUC (Δ AUC), Median (IQR) | P Value ^b |
|---|-----|--------------------------|-------------------------|-----------------------------|-------------------------|--|-------------------------|
| FRS adequately calculated and used | | | | | | | |
| Yes | 20 | 0.66 (0.59 to 0.73) | .001 | 0.70 (0.66 to 0.75) | .001 | 0.017 (−0.008 to 0.042) | .10 |
| No | 68 | 0.77 (0.72 to 0.83) | | 0.77 (0.72 to 0.82) | | 0.007 (−0.003 to 0.017) | |
| CHD outcome examined | | | | | | | |
| Yes | 57 | 0.77 (0.71 to 0.83) | .025 | 0.77 (0.73 to 0.82) | .12 | 0.006 (−0.004 to 0.016) | .12 |
| No | 31 | 0.70 (0.62 to 0.79) | | 0.72 (0.66 to 0.79) | | 0.010 (−0.015 to 0.035) | |
| Modeling of additional predictor may affect improvement ^c | | | | | | | |
| Yes | 20 | 0.64 (0.57 to 0.72) | <.001 | 0.68 (0.60 to 0.79) | .01 | 0.020 (0.000 to 0.040) | .004 |
| No | 68 | 0.77 (0.72 to 0.83) | | 0.77 (0.73 to 0.82) | | 0.005 (−0.002 to 0.014) | |
| Population as intended for FRS (<10% nonwhite and no CHD at baseline) | | | | | | | |
| Yes | 26 | 0.74 (0.65 to 0.83) | .51 | 0.75 (0.69 to 0.82) | .64 | 0.010 (0.000 to 0.020) | .98 |
| No | 62 | 0.74 (0.61 to 0.87) | | 0.75 (0.70 to 0.80) | | 0.007 (−0.003 to 0.017) | |
| Multivariable regression adequately documented | | | | | | | |
| Yes | 59 | 0.77 (0.71 to 0.83) | .001 | 0.78 (0.73 to 0.84) | .01 | 0.003 (−0.005 to 0.011) | <.001 |
| No | 29 | 0.69 (0.63 to 0.75) | | 0.72 (0.69 to 0.75) | | 0.016 (−0.004 to 0.036) | |
| AUC analysis adequately documented | | | | | | | |
| Yes | 45 | 0.77 (0.70 to 0.85) | .48 | 0.77 (0.72 to 0.83) | .74 | 0.002 (−0.008 to 0.012) | .009 |
| No | 43 | 0.72 (0.66 to 0.78) | | 0.74 (0.69 to 0.80) | | 0.010 (0.005 to 0.015) | |

Abbreviations: AD, additional predictor; AUC, area under the receiver operating curve; CHD, coronary heart disease; FRS, Framingham risk score; IQR, interquartile range.

^aStudies may contribute more than 1 analysis if they report more than 1 outcome or predictor or if results are presented in subgroups only. None of the studies reporting AUC analysis used only a subgroup of participants based on their FRS, and analysis of this feature was not performed. Only 7 studies also reported reclassification or calibration, and analyses of these features were not performed.

^bP value calculated from Mann-Whitney U test for the comparison of FRS vs FRS+AD groups.

^cAuthors examined different models of analysis (eg, continuous and categorical modeling of predictor or different splits in categorization) and presented the results for only 1 model without clarifying whether it was the best-fitting one.

also lower when the AUC was calculated for an outcome different from what the FRS was initially developed for. In such cases, the FRS is transformed into a “straw-man” predictive model and even mediocre new candidate risk factors can achieve an improvement. This result is analogous to observations from randomized clinical trials where drugs may show exaggerated treatment effects when the comparator drug is ineffective.⁹⁴

Furthermore, inadequate use of regression models and AUC analyses and modeling of additional predictors were associated with inflated estimates of improvement in the AUC. Studies without these methodological flaws did not show any substantial improvements in AUC. Publication bias and other types of selective reporting bias might also affect improved performance if articles and models that show no improvement in AUC remain unpublished. Misuse of subgroup analysis was also observed, as has been previously documented in diverse other fields.⁹⁰

We used a convenient systematic sampling strategy to select eligible studies. Many other articles might exam-

ine improved prediction of a risk factor beyond the FRS without citing the reference FRS article. However, it is unlikely that those studies would be better reported. Furthermore, reported methods are not always commensurate with what was actually done in a study,^{95,96} which is the case even more in predictive literature, where no widely accepted reporting standards are available. However, when the reported information suggests considerable problems or biases, it is unlikely the actual study was immune to them.¹⁴

Another limitation is that we considered only studies that had at least 1 analysis performed where incremental predictive ability was assessed. These articles are the most sophisticated in the large literature about studies of new candidate predictors. Numerous other articles only considered candidate risk factors for association with some outcome alone or simply juxtaposed them with the FRS in plain correlation evaluations without establishing incremental prediction. Empirical evidence from other fields, eg, cancer, suggests that almost all of these studies report significant associations for the examined pre-

dictors.⁴ Methodological problems in these studies are even more extensive.⁵

Our analyses focused on cardiovascular research. However, their implications probably extend to prediction models for other diseases. Rigorous design, analysis, and reporting are essential elements of research studies. Efforts to enhance the transparency of reporting of prognostic marker and predictive model studies⁶ require wider adoption. Currently, incremental predictive effects are difficult to discern from the spurious effects that biased analysis or reporting can create.

Author Affiliations: Department of Epidemiology and Public Health, Imperial College of Medicine, London, England (Dr Tzoulaki); Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece (Drs Liberopoulos and Ioannidis); Biomedical Research Institute, Foundation for Research and Technology–Hellas, Ioannina (Dr Ioannidis); and Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, and Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts (Dr Ioannidis).

Author Contributions: Dr Ioannidis had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Tzoulaki, Liberopoulos, Ioannidis.

Acquisition of data: Tzoulaki, Liberopoulos, Ioannidis.

Analysis and interpretation of data: Tzoulaki, Liberopoulos, Ioannidis.

Drafting of the manuscript: Tzoulaki, Ioannidis.

Critical revision of the manuscript for important intellectual content: Liberopoulos, Ioannidis.

Statistical analysis: Tzoulaki, Liberopoulos, Ioannidis.

Administrative, technical, or material support: Tzoulaki.

Study supervision: Ioannidis.

Financial Disclosures: None reported.

Additional Information: The eTables and eFigure are available at <http://www.jama.com>.

REFERENCES

- Allison M. Is personalized medicine finally arriving? *Nat Biotechnol*. 2008;26(5):509-517.
- Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat*. 1998;52(1-3):289-303.
- Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
- Kyzas PA, Denaxa-Kyza D, Ioannidis JP. Almost all articles on cancer prognostic markers report statistically significant results. *Eur J Cancer*. 2007;43(17):2559-2579.
- Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst*. 2005;97(14):1043-1055.
- McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst*. 2005;97(16):1180-1184.
- 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(18):1837-1847.
- 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*. 2002;106(25):3143-3421.
- Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935.
- Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker. *Stat Med*. 2008;27(2):157-172.
- Raggi P, Cooil B, Callister TQ. Use of electron beam tomography data to develop models for prediction of hard coronary events. *Am Heart J*. 2001;141(3):375-382.
- Meigs JB, Nathan DM, D'Agostino RB Sr, Wilson PW; Framingham Offspring Study. Fasting and post-challenge glycemia and cardiovascular disease risk. *Diabetes Care*. 2002;25(10):1845-1850.
- Kondos GT, Hoff JA, Sevrukov A, et al. Electron-beam tomography coronary artery calcium and cardiac events. *Circulation*. 2003;107(20):2571-2576.
- Folsom AR, Chambless LE, Duncan BB, Gilbert AC, Pankow JS; Atherosclerosis Risk in Communities Study Investigators. Prediction of coronary heart disease in middle-aged adults with diabetes. *Diabetes Care*. 2003;26(10):2777-2784.
- Raggi P, Shaw LJ, Berman DS, Callister TQ. Prognostic value of coronary artery calcium screening in subjects with and without diabetes. *J Am Coll Cardiol*. 2004;43(9):1663-1669.
- Lloyd-Jones DM, Nam BH, D'Agostino RB Sr, et al. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. *JAMA*. 2004;291(18):2204-2211.
- Pai JK, Pischon T, Ma J, et al. Inflammatory markers and the risk of coronary heart disease in men and women. *N Engl J Med*. 2004;351(25):2599-2610.
- Meisinger C, Loewel H, Mraz W, Koenig W. Prognostic value of apolipoprotein B and A-I in the prediction of myocardial infarction in middle-aged men and women. *Eur Heart J*. 2005;26(3):271-278.
- Ridker PM, Rifai N, Cook NR, Bradwin G, Buring JE. Non-HDL cholesterol, apolipoproteins A-I and B100, standard lipid measures, lipid ratios, and CRP as risk factors for cardiovascular disease in women. *JAMA*. 2005;294(3):326-333.
- Folsom AR, Chambless LE, Ballantyne CM, et al. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers. *Arch Intern Med*. 2006;166(13):1368-1373.
- Humphries SE, Cooper JA, Talmud PJ, Miller GJ. Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men. *Clin Chem*. 2007;53(1):8-16.
- Pischon T, Mohlig M, Hoffmann K, et al. Comparison of relative and attributable risk of myocardial infarction and stroke according to C-reactive protein and low-density lipoprotein cholesterol levels. *Eur J Epidemiol*. 2007;22(7):429-438.
- 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(7):776-785.
- 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(3):467-474.
- Detrano R, Guerci AD, Carr JJ, et al. Coronary calcium as a predictor of coronary events in four racial or ethnic groups. *N Engl J Med*. 2008;358(13):1336-1345.
- Wilson PW, Bozeman SR, Burton TM, Hoaglin DC, Ben-Joseph R, Pashos CL. Prediction of first events of coronary heart disease and stroke with consideration of adiposity. *Circulation*. 2008;118(2):124-130.
- Morrison AC, Bare LA, Luke MM, et al. Single nucleotide polymorphisms associated with coronary heart disease predict incident ischemic stroke in the atherosclerosis risk in communities study. *Cerebrovasc Dis*. 2008;26(4):420-424.
- Williams JE, Nieto FJ, Sanford CP, Couper DJ, Tyroler HA. The association between trait anger and incident stroke risk. *Stroke*. 2002;33(1):13-19.
- Gallo WT, Bradley EH, Falba TA, et al. Involuntary job loss as a risk factor for subsequent myocardial infarction and stroke. *Am J Ind Med*. 2004;45(5):408-416.
- Gallo WT, Teng HM, Falba TA, Kasl SV, Krumholz HM, Bradley EH. The impact of late career job loss on myocardial infarction and stroke. *Occup Environ Med*. 2006;63(10):683-687.
- Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med*. 2006;355(25):2631-2639.
- Jahnke C, Nagel E, Gebker R, et al. Prognostic value of cardiac magnetic resonance stress tests. *Circulation*. 2007;115(13):1769-1776.
- Price JF, Tzoulaki I, Lee AJ, Fowkes FG. Ankle brachial index and intima media thickness predict cardiovascular events similarly and increased prediction when combined. *J Clin Epidemiol*. 2007;60(10):1067-1075.
- Salameh MJ, Rundek T, Boden-Albala B, et al. Self-reported peripheral arterial disease predicts future vascular events in a community-based cohort. *J Gen Intern Med*. 2008;23(9):1423-1428.
- Zakai NA, Katz R, Jenny NS, et al. Inflammation and hemostasis biomarkers and cardiovascular risk in the elderly. *J Thromb Haemost*. 2007;5(6):1128-1135.
- Acevedo M, Pearce GL, Kottke-Marchant K, Sprecher DL. Elevated fibrinogen and homocysteine levels enhance the risk of mortality in patients from a high-risk preventive cardiology clinic. *Arterioscler Thromb Vasc Biol*. 2002;22(6):1042-1045.
- Brugger-Andersen T, Aarsetoy H, Grundt H, Staines H, Nilsen DW. The long-term prognostic value of multiple biomarkers following a myocardial infarction. *Thromb Res*. 2008;123(1):60-66.
- Tindle HA, Chang YF, Kuller LH, et al. Optimism, cynical hostility, and incident coronary heart disease and mortality in the Women's Health Initiative. *Circulation*. 2009;120(8):656-662.
- Arnlov J, Evans JC, Meigs JB, et al. Low-grade albuminuria and incidence of cardiovascular disease events in nonhypertensive and nondiabetic individuals. *Circulation*. 2005;112(7):969-975.
- Baldassarre D, Amato M, Pustina L, et al. Measurement of carotid artery intima-media thickness in dyslipidemic patients increases the power of traditional risk factors to predict cardiovascular events. *Atherosclerosis*. 2007;191(2):403-408.
- Mackness B, Durrington P, McElduff P, et al. Low paraoxonase activity predicts coronary events in the Caerphilly Prospective Study. *Circulation*. 2003;107(22):2775-2779.
- Hunt KJ, Sharrett AR, Chambless LE, Folsom AR, Evans GW, Heiss G. Acoustic shadowing on B-mode ultrasound of the carotid artery predicts CHD. *Ultrasound Med Biol*. 2001;27(3):357-365.
- Mora S, Redberg RF, Cui Y, et al. Ability of exercise testing to predict cardiovascular and all-cause death in asymptomatic women. *JAMA*. 2003;290(12):1600-1607.
- Sajadieh A, Nielsen OW, Rasmussen V, Hein HO, Hansen JF. Prevalence and prognostic significance of daily-life silent myocardial ischaemia in middle-aged and elderly subjects with no apparent heart disease. *Eur Heart J*. 2005;26(14):1402-1409.
- Cushman M, Arnold AM, Psaty BM, et al. C-reactive protein and the 10-year incidence of coronary heart disease in older men and women. *Circulation*. 2005;112(1):25-31.
- Dekker JM, Girman C, Rhodes T, et al. Metabolic syndrome and 10-year cardiovascular disease risk in the Hoorn Study. *Circulation*. 2005;112(5):666-673.
- Boekholdt SM, Hack CE, Sandhu MS, et al. C-reactive protein levels and coronary artery disease incidence and mortality in apparently healthy men and women. *Atherosclerosis*. 2006;187(2):415-422.
- Sajadieh A, Nielsen OW, Rasmussen V, et al. Ventricular arrhythmias and risk of death and acute myocardial infarction in apparently healthy subjects of age >or=55 years. *Am J Cardiol*. 2006;97(9):1351-1357.
- Hu G, Tuomilehto J, Borodulin K, Jousilahti P. The joint associations of occupational, commuting, and leisure-time physical activity, and the Framingham risk score on the 10-year risk of coronary heart disease. *Eur Heart J*. 2007;28(4):492-498.
- El Harchaoui K, van der Steeg WA, Stroes ES, et al. Value of low-density lipoprotein particle number and size as predictors of coronary artery disease in apparently healthy men and women. *J Am Coll Cardiol*. 2007;49(5):547-553.
- Arsenault BJ, Lemieux I, Despres JP, et al. Cholesterol levels in small LDL particles predict the risk of coronary heart disease in the EPIC-Norfolk prospective population study. *Eur Heart J*. 2007;28(22):2770-2777.
- Schousboe JT, Taylor BC, Kiel DP, Ensrud KE, Wilson KE, McCloskey EV. Abdominal aortic calcification detected on lateral spine images from a bone densitometer predicts incident myocardial infarction or stroke in older women. *J Bone Miner Res*. 2008;23(3):409-416.

53. Levitzky YS, Cupples LA, Murabito JM, et al. Prediction of intermittent claudication, ischemic stroke, and other cardiovascular disease by detection of abdominal aortic calcific deposits by plain lumbar radiographs. *Am J Cardiol*. 2008;101(3):326-331.
54. Fowkes FG, Murray GD, Butcher I, et al; Ankle Brachial Index Collaboration. Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality. *JAMA*. 2008;300(2):197-208.
55. Koerselman J, de Jaegere PP, Verhaar MC, Grobbee DE, van der Graaf Y; SMART Study Group. Prognostic significance of coronary collaterals in patients with coronary heart disease having percutaneous transluminal coronary angioplasty. *Am J Cardiol*. 2005;96(3):390-394.
56. Rundek T, Arif H, Boden-Albala B, Elkind MS, Paik MC, Sacco RL. Carotid plaque, a subclinical precursor of vascular events. *Neurology*. 2008;70(14):1200-1207.
57. Daniels LB, Laughlin GA, Clopton P, Maisel AS, Barrett-Connor E. Minimally elevated cardiac troponin T and elevated N-terminal pro-B-type natriuretic peptide predict mortality in older adults. *J Am Coll Cardiol*. 2008;52(6):450-459.
58. Chambless LE, Folsom AR, Sharrett AR, et al. Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study. *J Clin Epidemiol*. 2003;56(9):880-890.
59. Boutouyrie P, Tropeano AI, Asmar R, et al. Aortic stiffness is an independent predictor of primary coronary events in hypertensive patients. *Hypertension*. 2002;39(1):10-15.
60. Weiner DE, Tighiouart H, Griffith JL, et al. Kidney disease, Framingham risk scores, and cardiac and mortality outcomes. *Am J Med*. 2007;120(6):552-558.
61. 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(1):21-29.
62. Ridker PM, Rifai N, Rose L, Buring JE, Cook NR. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *N Engl J Med*. 2002;347(20):1557-1565.
63. Vliegenthart R, Oudkerk M, Hofman A, et al. Coronary calcification improves cardiovascular risk prediction in the elderly. *Circulation*. 2005;112(4):572-577.
64. van der Steeg WA, Boekholdt SM, Stein EA, et al. Role of the apolipoprotein B-apolipoprotein A-I ratio in cardiovascular risk assessment. *Ann Intern Med*. 2007;146(9):640-648.
65. Meuwese MC, Stroes ES, Hazen SL, et al. Serum myeloperoxidase levels are associated with the future risk of coronary artery disease in apparently healthy individuals. *J Am Coll Cardiol*. 2007;50(2):159-165.
66. Cao JJ, Biggs ML, Barzilay J, et al. Cardiovascular and mortality risk prediction and stratification using urinary albumin excretion in older adults ages 68-102. *Atherosclerosis*. 2008;197(2):806-813.
67. Shah T, Casas JP, Cooper JA, et al. Critical appraisal of CRP measurement for the prediction of coronary heart disease events. *Int J Epidemiol*. 2009;38(1):217-231.
68. Mainous AG III, Everett CJ, Player MS, King DE, Diaz VA. Importance of a patient's personal health history on assessments of future risk of coronary heart disease. *J Am Board Fam Med*. 2008;21(5):408-413.
69. Balady GJ, Larson MG, Vasan RS, Leip EP, O'Donnell CJ, Levy D. Usefulness of exercise testing in the prediction of coronary disease risk among asymptomatic persons as a function of the Framingham risk score. *Circulation*. 2004;110(14):1920-1925.
70. Denes P, Larson JC, Lloyd-Jones DM, Prineas RJ, Greenland P. Major and minor ECG abnormalities in asymptomatic women and risk of cardiovascular events and mortality. *JAMA*. 2007;297(9):978-985.
71. Wilson PW, Nam BH, Pencina M, D'Agostino RB Sr, Benjamin EJ, O'Donnell CJ. C-reactive protein and risk of cardiovascular disease in men and women from the Framingham Heart Study. *Arch Intern Med*. 2005;165(21):2473-2478.
72. Mallat Z, Benessiano J, Simon T, et al. Circulating secretory phospholipase A2 activity and risk of incident coronary events in healthy men and women. *Arterioscler Thromb Vasc Biol*. 2007;27(5):1177-1183.
73. Greenland P, LaBree L, Azen SP, Doherty TM, Detrano RC. Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals. *JAMA*. 2004;291(2):210-215.
74. Koenig W, Lowel H, Baumert J, Meisinger C. C-reactive protein modulates risk prediction based on the Framingham Score: implications for future risk assessment. *Circulation*. 2004;109(11):1349-1353.
75. Mora S, Yanek LR, Moy TF, Fallin MD, Becker LC, Becker DM. Interaction of body mass index and Framingham risk score in predicting incident coronary disease in families. *Circulation*. 2005;111(15):1871-1876.
76. Simmons RK, Sharp S, Boekholdt SM, et al. Evaluation of the Framingham risk score in the European Prospective Investigation of Cancer-Norfolk cohort. *Arch Intern Med*. 2008;168(11):1209-1216.
77. Lee J, Heng D, Ma S, Chew SK, Hughes K, Tai ES. The metabolic syndrome and mortality. *Clin Endocrinol (Oxf)*. 2008;69(2):225-230.
78. Strom Moller C, Zethelius B, Sundstrom J, Lind L. Persistent ischaemic ECG abnormalities on repeated ECG examination have important prognostic value for cardiovascular disease beyond established risk factors. *Heart*. 2007;93(9):1104-1110.
79. Hassinen M, Komulainen P, Lakka TA, et al. Metabolic syndrome and the progression of carotid intima-media thickness in elderly women. *Arch Intern Med*. 2006;166(4):444-449.
80. Jeppesen J, Hansen TW, Rasmussen S, Ibsen H, Torp-Pedersen C, Madsbad S. Insulin resistance, the metabolic syndrome, and risk of incident cardiovascular disease. *J Am Coll Cardiol*. 2007;49(21):2112-2119.
81. Gulati M, Arnsdorf MF, Shaw LJ, et al. Prognostic value of the duke treadmill score in asymptomatic women. *Am J Cardiol*. 2005;96(3):369-375.
82. Corona AJ, Martinez DR, Avila MH, et al. Microalbuminuria as a predictor of myocardial infarction in a Mexican population. *Kidney Int Suppl*. 2005;97(suppl):S34-S39.
83. Hansen TW, Staessen JA, Torp-Pedersen C, et al. Ambulatory arterial stiffness index predicts stroke in a general population. *J Hypertens*. 2006;24(11):2247-2253.
84. Olsen MH, Wachtell K, Ibsen H, et al; LIFE Study Investigators. Reductions in albuminuria and in electrocardiographic left ventricular hypertrophy independently improve prognosis in hypertension. *J Hypertens*. 2006;24(4):775-781.
85. Stork S, Feelders RA, van den Beld AW, et al. Prediction of mortality risk in the elderly. *Am J Med*. 2006;119(6):519-525.
86. Schouten BW, Bohnen AM, Bosch JL, et al. Erectile dysfunction prospectively associated with cardiovascular disease in the Dutch general population. *Int J Impot Res*. 2008;20(1):92-99.
87. Nawrot TS, Staessen JA, Thijs L, et al. Should pulse pressure become part of the Framingham risk score? *J Hum Hypertens*. 2004;18(4):279-286.
88. Shaw LJ, Raggi P, Schisterman E, Berman DS, Callister TQ. Prognostic value of cardiac risk factors and coronary artery calcium screening for all-cause mortality. *Radiology*. 2003;228(3):826-833.
89. Stern MP, Williams K, Gonzalez-Villalpando C, Hunt KJ, Haffner SM. Does the metabolic syndrome improve identification of individuals at risk of type 2 diabetes and/or cardiovascular disease? *Diabetes Care*. 2004;27(11):2676-2681.
90. Patsopoulos NA, Tatsioni A, Ioannidis JP. Claims of sex differences. *JAMA*. 2007;298(8):880-893.
91. Hlatky MA, Greenland P, Arnett DK, et al; American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council. Criteria for evaluation of novel markers of cardiovascular risk. *Circulation*. 2009;119(17):2408-2416.
92. Ioannidis JP. Effect of formal statistical significance on the credibility of observational associations. *Am J Epidemiol*. 2008;168(4):374-383.
93. McGeehan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice. *Arch Intern Med*. 2008;168(21):2304-2310.
94. Bero L, Oostvogel F, Bacchetti P, Lee K. Factors associated with findings of published trials of drug-drug comparisons. *PLoS Med*. 2007;4(6):e184.
95. Ioannidis JP, Lau J. Can quality of clinical trials and meta-analyses be quantified? *Lancet*. 1998;352(9128):590-591.
96. Schulz KF, Grimes DA. Allocation concealment in randomised trials. *Lancet*. 2002;359(9306):614-618.