Comparisons of established risk prediction models for cardiovascular disease: systematic review

George C M Siontis research associate¹, Ioanna Tzoulaki lecturer¹, Konstantinos C Siontis research associate¹, John P A Ioannidis professor²

¹Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece; ²Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305-5411, USA

Abstract

Objective To evaluate the evidence on comparisons of established cardiovascular risk prediction models and to collect comparative information on their relative prognostic performance.

Design Systematic review of comparative predictive model studies.

Data sources Medline and screening of citations and references.

Study selection Studies examining the relative prognostic performance of at least two major risk models for cardiovascular disease in general populations.

Data extraction Information on study design, assessed risk models, and outcomes. We examined the relative performance of the models (discrimination, calibration, and reclassification) and the potential for outcome selection and optimism biases favouring newly introduced models and models developed by the authors.

Results 20 articles including 56 pairwise comparisons of eight models (two variants of the Framingham risk score, the assessing cardiovascular risk to Scottish Intercollegiate Guidelines Network to assign preventative treatment (ASSIGN) score, systematic coronary risk evaluation (SCORE) score, Prospective Cardiovascular Münster (PROCAM) score, QRESEARCH cardiovascular risk (QRISK1 and QRISK2) algorithms, Reynolds risk score) were eligible. Only 10 of 56 comparisons exceeded a 5% relative difference based on the area under the receiver operating characteristic curve. Use of other discrimination, calibration, and reclassification statistics was less consistent. In 32 comparisons, an outcome was used that had been used in the original development of only one of the compared models, and in 25 of these comparisons (78%) the outcome-congruent model had a better area under the receiver operating characteristic curve. Moreover, authors always reported better area under the receiver operating characteristic curves for models that they themselves developed (in five articles on newly introduced models and in three articles on subsequent evaluations).

Conclusions Several risk prediction models for cardiovascular disease are available and their head to head comparisons would benefit from standardised reporting and formal, consistent statistical comparisons. Outcome selection and optimism biases apparently affect this literature.

Introduction

Cardiovascular disease carries major morbidity and mortality.¹ To effectively implement prevention strategies clinicians need reliable tools to identify individuals without known cardiovascular disease who are at high risk of a cardiovascular event.² For this purpose, multivariable risk assessment tools, such as the Framingham risk score, are recommended for clinical use.³ Besides the Framingham risk score, several other risk prediction tools combining different sets of variables have been developed and validated.⁴ Some investigators have evaluated the performance of two or more risk prediction models in the same populations.

We evaluated the evidence on comparisons of established cardiovascular risk prediction models. We systematically collected comparative information on discrimination, calibration, and reclassification performance and evaluated whether specific biases may have affected the inferences of studies comparing such models.
Methods

Eligible models and literature search

We assessed prediction models for the risk of cardiovascular disease in general populations that were considered in two recent expert reviews: the Framingham risk score (and the national cholesterol education program–adult treatment panel III version), the assessing cardiovascular risk to Scottish Intercollegiate Guidelines Network to assign preventative treatment (ASSIGN) score, systematic coronary risk evaluation (SCORE) score, Prospective Cardiovascular Münster (PROCAM) score, QRESEARCH cardiovascular risk (QRISK1 and QRISK2) algorithms, Reynolds risk score, and the World Health Organization/International Society of Hypertension score. Different versions of the Framingham risk score were categorised as Framingham risk score (including the Framingham risk score described by Anderson et al for risk of coronary heart disease and stroke and the Framingham risk score proposed by Wilson et al) (also proposed by National Institute for Health and Clinical Excellence guidelines) and as FRS (CVD) (which included the global Framingham risk score equations to predict cardiovascular disease). See supplementary table 1 for additional details.

Medline (last update July 2011) was searched for articles with data on the performance of at least two of these models. We also scrutinised the received citations (through SCOPUS) and the references of all eligible papers for any additional relevant studies (see appendix for primary screening algorithm). Titles and abstracts were screened first and potentially eligible articles scrutinised in full text. No year or language restrictions were applied.

Study eligibility

Articles were eligible if they examined at least two pertinent risk models for the prediction of cardiovascular disease in populations without cardiovascular disease or general populations. We included original articles irrespective of sample size and duration of follow-up. Eligible outcomes were cardiovascular disease (and any composite cardiovascular disease end point), cardiovascular disease mortality, and coronary heart disease, including stable disease and acute coronary syndromes. When different published data on identical comparisons were identified comparing the same models, in the same cohort, and for the same outcome, we kept only the data that included the largest number of events. We excluded cross sectional studies, studies where all cause mortality was the only outcome, studies that used models to calculate the baseline risk without providing outcome data, and studies including exclusively patients with specific morbidities—that is, patients with known cardiovascular disease, diabetes, or other diseases.

Two investigators (GCMS, KCS) independently carried out the literature searches and assessed the studies for eligibility. Discrepancies were resolved by consensus and arbitration by two other investigators (IT, JPAl).

Data extraction

Two investigators independently extracted data from the main paper (GCMS, IT) and any accompanying supplemental material. The following items of interest were recorded in standardised forms: study design (prospective or retrospective), year of publication, sample size, type of population, percentage of baseline population with pre-existing cardiovascular disease, and reported risk models. We recorded the clinical end points assessed in each study (cardiovascular disease, cardiovascular disease mortality, coronary heart disease) and the respective number of events. When multiple different eligible outcomes or populations were identified in the same model comparison, we considered each outcome or cohort separately. Similarly, when more than two prognostic models were presented in an article, we considered all possible pairwise comparisons as eligible. Whenever a study also examined subgroups, such as males and females, we focused on the whole population unless only data per subgroup were provided; in those cases, we extracted data for each eligible subgroup separately.

Moreover, for each study we also captured whether the authors reported the presence of missing data on examined outcomes and on variables included in risk prediction models; and, if so, we recorded how missing data were managed (with imputation and by which methods, exclusion of missing observations, or other). We further extracted information on the geographical origin of each study and noted whether it was the same country to the one in which one (or both) of the compared models was initially developed.

For each model in each article we extracted metrics on discrimination (area under the receiver operating characteristic curve (or the equivalent C statistic), D statistic, R statistic, and Brier score), their 95% confidence intervals, and the P value for comparison between models when available. We also captured calibration and reclassification metrics. We extracted information on whether the observed versus predicted ratio and lack of fit statistics were reported, and whether the calibration plot was shown. Finally, we extracted information on reclassification statistics, such as the net reclassification index, and on the classification percentages of each model along with the thresholds used by each study.

Data analysis and evaluation of biases

We analysed each risk model pairwise comparison separately. For each comparison we noted the model with a numerically higher area under the receiver operating characteristic curve estimate, and whether there was formal statistical testing of the difference in areas under the receiver operating characteristic curve. When confidence intervals were not available, we estimated them as previously proposed. We also recorded separately which pairwise comparisons had a relative difference in area under the receiver operating characteristic curve exceeding 5% (for example, if the worse score had an area under the receiver operating characteristic curve of 0.70, the better score had one >0.70×1.05=0.735). The choice of a 5% threshold was chosen for descriptive purposes only. Furthermore, we noted whether models differed in other performance metrics. Calibration was considered better when the observed to predicted ratio was closer to 1.

We also evaluated the potential for outcome selection and optimism biases. Some of the examined risk scores have been originally developed for different cardiovascular outcomes (see supplementary table 1). We evaluated whether the examined outcome in each comparison was used in the original development of only one of the two compared models and, if so, whether the outcome-congruent model showed better performance. Owing to optimism bias, a new model may have better performance than the competing standard model when it is first presented, but not in subsequent comparisons. Therefore we noted whether each article described the application of previously established models or was the first to describe or validate a specific model or models. Moreover, authors who developed one model may favour publishing results that show its superiority against competing models. We thus noted whether
any of the study authors had been involved in the development of any of the assessed models. Finally, we recorded the authors’ comments on the relative performance of the model and examined whether these were affected by such potential biases. Analyses were done in Stata 10.1 (College Station, TX). P values are two tailed.

Results

Inclusion of studies

Of 672 published articles screened at title and abstract level, 74 were identified as potentially eligible for inclusion in the review. Of these, 58 articles were excluded because they only compared models using a baseline risk calculation without association with outcomes (n=20); assessed only patients with specific conditions (diabetes (n=11), HIV infection (n=4), known cardiovascular disease (n=3), liver transplantation (n=1), schizoaffective disorder (n=1), systemic lupus erythematosus or rheumatoid arthritis (n=1)); or had ineligible model comparisons (n=10), ineligible outcomes (non-cardiovascular disease outcomes) (n=6), or duplicate comparisons (n=1). (See supplementary web figure). Searches of references and citations yielded another four eligible articles. Overall, 20 articles were analysed (table 1). Of these, 58 articles were excluded because they only compared models using a baseline risk calculation without association with outcomes (n=20); assessed only patients with specific conditions (diabetes (n=11), HIV infection (n=4), known cardiovascular disease (n=3), liver transplantation (n=1), schizoaffective disorder (n=1), systemic lupus erythematosus or rheumatoid arthritis (n=1)); or had ineligible model comparisons (n=10), ineligible outcomes (non-cardiovascular disease outcomes) (n=6), or duplicate comparisons (n=1). (See supplementary web figure). Searches of references and citations yielded another four eligible articles. Overall, 20 articles were analysed (table 1). Of 28 possible types of pairwise comparisons of these eight risk scores, 14 existed in the literature. After excluding overlapping data (same models compared, same outcome, same cohort), independent data were available on 56 individual comparisons of risk models. Eight articles reported data for men and women separately (44 comparisons), four reported overall data (four comparisons), seven assessed only males (seven comparisons), and one assessed only women (one comparison, table 2). The Framingham risk score or FRS(CVD) were involved in 50 of 56 comparisons (tables 1 and 2). In four articles (eight comparisons) the authors reported information on missing data on the examined outcomes, and in all cases the investigators excluded the respective participants (see supplementary table 2). Information on missing data for variables included in risk models was reported in 11 articles (44 comparisons). Different strategies were implemented to deal with missing data and sometimes different strategies were applied to different predictors: exclusion of participants with missing data (27 comparisons), multiple imputation technique (16 comparisons), value generation by multivariate regression methods (10 comparisons), replacement by the mean value of the variable (nine comparisons), and assumption that participants without information on smoking were non-smokers (eight comparisons, also see supplementary table 2). In 25 comparisons, the geographical origin of the study population was the same as the origin of the population in which at least one of the examined models was initially developed (see supplementary table 3).

Characteristics of eligible studies and risk models

All articles were published after 2002 (table 1). All but two studies had prospective designs. Most (n=17) articles assessed populations of European descent. The median sample size was 8958 (interquartile range 2365-327136).

Eight different risk models were evaluated (all of those considered upfront eligible, except the World Health Organization/International Society of Hypertension score). Of the 28 possible types of pairwise comparisons of these eight risk scores, 14 existed in the literature. After excluding overlapping data (same models compared, same outcome, same cohort), independent data were available on 56 individual comparisons of risk models. Eight articles reported data for men and women separately (44 comparisons), four reported overall data (four comparisons), seven assessed only males (seven comparisons), and one assessed only women (one comparison, table 2). The Framingham risk score or FRS(CVD) were involved in 50 of 56 comparisons (tables 1 and 2). In four articles (eight comparisons) the authors reported information on missing data on the examined outcomes, and in all cases the investigators excluded the respective participants (see supplementary table 2). Information on missing data for variables included in risk models was reported in 11 articles (44 comparisons). Different strategies were implemented to deal with missing data and sometimes different strategies were applied to different predictors: exclusion of participants with missing data (27 comparisons), multiple imputation technique (16 comparisons), value generation by multivariate regression methods (10 comparisons), replacement by the mean value of the variable (nine comparisons), and assumption that participants without information on smoking were non-smokers (eight comparisons, also see supplementary table 2). In 25 comparisons, the geographical origin of the study population was the same as the origin of the population in which at least one of the examined models was initially developed (see supplementary table 3).

Discrimination performance

Area under the receiver operating characteristic curve estimates were available for all 56 pairwise comparisons (table 2). Confidence intervals were given for only 20 pairs and P values for the comparison of area under the receiver operating characteristic curve were available for only two comparisons (in a single study").

The relative difference between the area under the receiver operating characteristic curve estimates exceeded 5% in only 10 (18%) comparisons, but even these differences were inconsistent: compared with SCORE, the Framingham risk score was worse in two cases but better in another two; compared with PROCAM, the Framingham risk score was worse in one case but better in another three; finally, FRS(CVD) was worse than SCORE in two cases.

Among the 50 comparisons that included variants of the Framingham risk score, in 37 (74%) the area under the receiver operating characteristic curve estimate was higher for the comparator model.

Use of other discrimination metrics (D statistic, R² statistic, Brier score) was inconsistent. At least one of these metrics was available for 26 comparisons (see supplementary table 4).

Calibration

Calibration performance was reported in 38 comparisons (see supplementary table 5). Observed versus predicted ratio estimates were available for 23 comparisons and results were quite inconsistent. The Framingham risk score was better than FRS(CVD) in one comparison but worse in another. The Framingham risk score was worse than ASSIGN in two comparisons, SCORE in two, QRISK1 in five, and PROCAM in one comparison, but it was better than ASSIGN in two comparisons, PROCAM in two, and QRISK1 in one comparison. FRS(CVD) was worse than ASSIGN in two comparisons and QRISK1 in one comparison, but it was better than QRISK1 in another comparison. Finally, QRISK1 was better than ASSIGN in two comparisons.

The 95% confidence intervals of the observed to predicted ratio were available in only two comparisons, so we could not tell whether differences were beyond chance.

Risk reclassification

Reporting of risk classification and reclassification was uncommon; information was available for 10 comparisons. In nine comparisons a dichotomous cut-off point of 20% 10 year risk was used; one study used 0-5, 5-10, 10-20, >20% as risk thresholds. All comparisons reported the number of participants reclassified with use of alternative models along with the predicted and observed risk in each risk category. The net reclassification index was calculated for six comparisons between non-nested models, all using the 20% threshold: ASSIGN versus Framingham risk score (n=2, net reclassification index 4%, 16%); ASSIGN versus FRS(CVD) (n=2, 0%, 12%); and FRS(CVD) versus Framingham risk score (n=2, 4% for both).

Outcome selection bias

In 13 comparisons the examined outcome was the one for which both compared models had been developed and validated, whereas in 32 comparisons only one of the compared models had been originally developed for that outcome, and in the other 11 comparisons none of the compared models had been developed originally for that outcome. When an outcome was
used that had been used in the original development of only one of the compared models, it was more common for the outcome-congruent model to have a better area under the receiver operating characteristic curve than the comparator (25 vs 7, P<0.001, based on point estimates).

Optimism bias

Five articles (12 comparisons) described a model for the first time (table 3). In all 12 comparisons, the new model had a higher area under the receiver operating characteristic curve estimate than Framingham risk score versions, although the relative improvement exceeded 5% only for one model (PROCAM better than Framingham risk score). Ten subsequently published articles addressed one or more of these same comparisons (table 3). In three (13-15) articles at least one of the authors had been previously involved in the development of one of the compared models, and that model continued to have a better area under the receiver operating characteristic curve. Conversely, two of the seven articles (26 28 35 36-39) published by entirely independent authors showed the older model to have a better area under the receiver operating characteristic curve.

Author interpretation

Overall, the authors claimed superiority of one model in 31 of 56 comparisons (see supplementary table 3). In 25 of these 31 comparisons a Framingham risk score version was one of the models compared and in all 25 cases the comparator model was claimed to be superior. SCORE>Framingham risk score (n=3), ASSIGN>Framingham risk score (n=6), PROCAM>Framingham risk score (n=1), QRISK1>Framingham risk score (n=4), QRISK2>Framingham risk score (n=4), FRS (CVD)>Framingham risk score (n=2), ASSIGN>FRS (CVD) (n=2), QRISK1>FRS (CVD) (n=2), and Reynolds risk score>Framingham risk score (n=1). The other six pairs where superiority was claimed were QRISK2>QRISK1 (n=4) and QRISK1>ASSIGN (n=2). For 22 comparisons the authors either claimed that both models had good or equal discriminatory ability or did not comment on their relative performance. In eight articles the authors favoured models they had themselves developed (five first publications, three subsequent publications). Authors involved in the development of a model never favoured a comparator.

Discussion

Comparative studies on the relative performance of established risk models for prediction of cardiovascular disease often suggest that one model may be better than another. In particular, the Framingham risk score usually had inferior performance compared with other models, but the results were sometimes inconsistent across studies, and inferences may be susceptible to potential biases and methodological shortcomings. Most studies did not compare statistically the models that they examined. Models were usually reported to be superior against comparators when the examined outcome was the one that the model was developed for but not the one for which the comparator was developed. Articles presenting new models or including authors involved in the original development of a model favoured the model that the authors had developed.

Comparison with other studies

Head to head comparisons of emerging risk models are important to perform so as to document improvements in risk prediction. We showed that such data are limited and, when available, difficult to interpret. Discrimination, the ability of a statistical model to distinguish those who experience cardiovascular disease events from those who do not, was presented for all comparisons but the differences were usually small. Only in 18% of the comparisons did the relative difference between the two areas under the receiver operating characteristic curve exceed 5%. Most studies did not report the confidence intervals of the area under the receiver operating characteristic curve or the P values for the comparison between models. Calibration, which assesses how closely predicted estimates of absolute risk agree with actual outcomes, was reported in two thirds of the comparisons, but again formal statistical testing was lacking. Although the area under the receiver operating characteristic curve is the most commonly used discrimination metric, it has limitations.40 Similarly, assessment of model calibration by the Hosmer-Lemeshow goodness of fit test is sensitive to sample size and gives no information on the extent or direction of miscalibration.41 42 Evaluating calibration graphically either by 10ths of predicted risk or by key prognostic variables, such as age, is more informative than a single P value.

Assessment of risk reclassification was sparse and, when assessed, it was suboptimally described, in agreement with previous empirical evaluations.43 44 Reclassification is a clinically useful concept. It makes most sense when the categories of risk are clearly linked to different indications for interventions. It may be informative to report the percentage of patients changing risk categories and their direction of change. However, summary metrics such as the net reclassification index are problematic, especially when the compared models are non-nested (that is, they include different predictors and are derived from different datasets), and the problems are even worse when at least one model is poorly calibrated.45

Choices of comparators and outcomes are particularly important in such studies. Models were often claimed to be superior when the outcome examined was different from what the comparator model had been developed for. In those cases, the comparator is disadvantaged and becomes a strawman comparator towards which superiority can easily be claimed; a phenomenon analogous to that observed in clinical trial studies where an intervention is compared against a placebo or ineffective intervention.46 In addition, we observed some evidence of potential optimism bias, with potentially unwarranted belief in the predictive performance of newer models by the scientists developing them. Authors consistently claimed superiority of the models that they have developed versus comparators. While genuine progress in predictive ability is a possible explanation for this pattern, it is worthwhile to ensure that such favourable results are also validated by completely independent investigators.

Limitations of the study

Our study has limitations. Firstly, most of the analysed studies and models pertained to populations of European descent. Risk models may, however, perform differently in populations of different racial or ethnic backgrounds.47 48 Systematic efforts for model validation in other populations are essential.49 Secondly, most confidence intervals of area under the receiver operating characteristic curve estimates were unavailable and were derived as previously described.50 We examined whether 95% confidence intervals did or did not overlap. A more formal statistical testing would have required access to individual level data to account for the fact that models were evaluated in the same population in each comparison using the pairwise individual level correlation in the calculations.51
Conclusions

Current studies comparing predictive models often have limitations or are missing information, which makes it difficult to reach robust conclusions about the best model or the ranking of performance of models. It should also be acknowledged that the answers to these questions may be different in different populations and settings. The box shows several items and pieces of information that would be useful to consider in the design and reporting of results in studies comparing different predictive models, to make these evaluations more useful, unbiased, and transparent, and to allow a balanced interpretation of the relative performance of these models.

The clinical usefulness of these models should be ultimately established on the basis of their potential for affecting decisions on treatment and prevention and improving health outcomes. Ideally, this would require randomised trials where patients are allocated to being managed using information from different predictive models. Given that such trials are difficult to perform and costly, evidence from well conducted studies of comparative predictive performance will remain important. Our empirical evaluation suggests that such studies may benefit from using standardised bootstrapping, calibration, and reclassification metrics with formal statistical comparisons; and standardised outcomes that are clinically appropriate and, whenever possible, relevant to both compared models. Finally, improved performance of new models versus established ones should ideally be documented in several studies carried out by independent investigators.

Contributors: GCMS, IT, KCS, and JPAI conceived the study, analysed the data, interpreted the results, and drafted the manuscript. GCMS and IT extracted the data. JPAI is the guarantor.

Funding: This study received no additional funding.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support for any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; and no other relationships or activities that could have influenced the submitted work.

Ethical approval: Not required.

Funding: This study received no additional funding.

Ethical approval: Not required.

The clinical usefulness of these models should be ultimately established on the basis of their potential for affecting decisions on treatment and prevention and improving health outcomes. Ideally, this would require randomised trials where patients are allocated to being managed using information from different predictive models. Given that such trials are difficult to perform and costly, evidence from well conducted studies of comparative predictive performance will remain important. Our empirical evaluation suggests that such studies may benefit from using standardised bootstrapping, calibration, and reclassification metrics with formal statistical comparisons; and standardised outcomes that are clinically appropriate and, whenever possible, relevant to both compared models. Finally, improved performance of new models versus established ones should ideally be documented in several studies carried out by independent investigators.

Conclusions
Current studies comparing predictive models often have limitations or are missing information, which makes it difficult to reach robust conclusions about the best model or the ranking of performance of models. It should also be acknowledged that the answers to these questions may be different in different populations and settings. The box shows several items and pieces of information that would be useful to consider in the design and reporting of results in studies comparing different predictive models, to make these evaluations more useful, unbiased, and transparent, and to allow a balanced interpretation of the relative performance of these models.

The clinical usefulness of these models should be ultimately established on the basis of their potential for affecting decisions on treatment and prevention and improving health outcomes. Ideally, this would require randomised trials where patients are allocated to being managed using information from different predictive models. Given that such trials are difficult to perform and costly, evidence from well conducted studies of comparative predictive performance will remain important. Our empirical evaluation suggests that such studies may benefit from using standardised bootstrapping, calibration, and reclassification metrics with formal statistical comparisons; and standardised outcomes that are clinically appropriate and, whenever possible, relevant to both compared models. Finally, improved performance of new models versus established ones should ideally be documented in several studies carried out by independent investigators.

Conclusions

Current studies comparing predictive models often have limitations or are missing information, which makes it difficult to reach robust conclusions about the best model or the ranking of performance of models. It should also be acknowledged that the answers to these questions may be different in different populations and settings. The box shows several items and pieces of information that would be useful to consider in the design and reporting of results in studies comparing different predictive models, to make these evaluations more useful, unbiased, and transparent, and to allow a balanced interpretation of the relative performance of these models.

The clinical usefulness of these models should be ultimately established on the basis of their potential for affecting decisions on treatment and prevention and improving health outcomes. Ideally, this would require randomised trials where patients are allocated to being managed using information from different predictive models. Given that such trials are difficult to perform and costly, evidence from well conducted studies of comparative predictive performance will remain important. Our empirical evaluation suggests that such studies may benefit from using standardised bootstrapping, calibration, and reclassification metrics with formal statistical comparisons; and standardised outcomes that are clinically appropriate and, whenever possible, relevant to both compared models. Finally, improved performance of new models versus established ones should ideally be documented in several studies carried out by independent investigators.

Conclusions

Current studies comparing predictive models often have limitations or are missing information, which makes it difficult to reach robust conclusions about the best model or the ranking of performance of models. It should also be acknowledged that the answers to these questions may be different in different populations and settings. The box shows several items and pieces of information that would be useful to consider in the design and reporting of results in studies comparing different predictive models, to make these evaluations more useful, unbiased, and transparent, and to allow a balanced interpretation of the relative performance of these models.

The clinical usefulness of these models should be ultimately established on the basis of their potential for affecting decisions on treatment and prevention and improving health outcomes. Ideally, this would require randomised trials where patients are allocated to being managed using information from different predictive models. Given that such trials are difficult to perform and costly, evidence from well conducted studies of comparative predictive performance will remain important. Our empirical evaluation suggests that such studies may benefit from using standardised bootstrapping, calibration, and reclassification metrics with formal statistical comparisons; and standardised outcomes that are clinically appropriate and, whenever possible, relevant to both compared models. Finally, improved performance of new models versus established ones should ideally be documented in several studies carried out by independent investigators.

Conclusions

Current studies comparing predictive models often have limitations or are missing information, which makes it difficult to reach robust conclusions about the best model or the ranking of performance of models. It should also be acknowledged that the answers to these questions may be different in different populations and settings. The box shows several items and pieces of information that would be useful to consider in the design and reporting of results in studies comparing different predictive models, to make these evaluations more useful, unbiased, and transparent, and to allow a balanced interpretation of the relative performance of these models.

The clinical usefulness of these models should be ultimately established on the basis of their potential for affecting decisions on treatment and prevention and improving health outcomes. Ideally, this would require randomised trials where patients are allocated to being managed using information from different predictive models. Given that such trials are difficult to perform and costly, evidence from well conducted studies of comparative predictive performance will remain important. Our empirical evaluation suggests that such studies may benefit from using standardised bootstrapping, calibration, and reclassification metrics with formal statistical comparisons; and standardised outcomes that are clinically appropriate and, whenever possible, relevant to both compared models. Finally, improved performance of new models versus established ones should ideally be documented in several studies carried out by independent investigators.
Suggestions for studies comparing risk prediction models

• Comparative studies should be carried out in independent samples from those where each model was originally developed, and ideally by investigators other than those who originally proposed these models.

• The study setting, country, and type of population should be described; it should also be recognised whether these characteristics are expected to offer any clear advantage to one of the compared models.

• The main outcome of the study should be clearly defined and clinically relevant; it should be recognised that models originally developed to predict other outcomes may exhibit inferior predictive performance.

• Models should be calculated using the same exact predictors and coefficients as when they were originally developed and validated.

• The follow-up time should correspond to the same follow-up as when the models were developed (for example, 10 year risk); deviations should be clarified and an explanation about choice given.

• The discrimination of each model should be given with point estimates and confidence intervals; differences between the discrimination of compared models should be formally tested, reporting the magnitude of the difference and the accompanying uncertainty.

• The calibration of each model may be assessed with statistical tests, but there is no good formal test for comparing calibration performance; it is useful to also show graphically the expected versus predicted risk for different levels of risk or levels of predictors.

• Examination of reclassification performance of examined risk scores is meaningful when there are well established clinically relevant risk thresholds, it is useful to provide information on the number of correct and incorrect classifications; avoid using the net reclassification improvement for non-nested models.

• The extent of missing information for outcomes and predictors should be described, also explaining how missing information was handled.

What is already known on this topic

Several risk prediction models for cardiovascular disease are recommended for clinical use; these models have often been developed and validated in different populations and for different outcomes.

The comparative prognostic performance of the most popular and widely used risk models in terms of discrimination, calibration, and reclassification is largely unknown.

What this study adds

Data from 20 studies (56 model comparisons) show limited evidence and inconsistent results about the relative prognostic ability of the most popular risk prediction models for cardiovascular disease.

The literature seems to be affected by optimism and outcome selection biases.

Standardised methodology and reporting could improve the quality of comparative studies of predictive models and guide future efforts towards meaningful prognostic research.

45 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. New approaches to evaluating predictive accuracy. Stat Med 2008;27:2665-70.
46 Ioannidis JP, Truelsen T. Effect of new evidence on reporting of net reclassification improvement from updated meta-analyses. JAMA 2009;301:2591-7.
47 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:83-69.
48 Ioannidis JP, Truelsen T. Effect of new evidence on reporting of net reclassification improvement from updated meta-analyses. JAMA 2009;301:2591-7.
49 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:83-69.
50 Chu X, D'Agostino RB Sr, Vasan RS. How to interpret a net reclassification improvement analysis. J Clin Epidemiol 2009;62:1242-50.
51 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:83-69.
52 Ioannidis JP, Truelsen T. Effect of new evidence on reporting of net reclassification improvement from updated meta-analyses. JAMA 2009;301:2591-7.

Accepted: 6 April 2012

Cite this as: BMJ 2012;344:e3318

This is an open-access article distributed under the terms of the Creative Commons Attribution Non-commercial License, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited, the use is non commercial and is otherwise in compliance with the license. See: http://creativecommons.org/licenses/by-nc/2.0/ and http://creativecommons.org/licenses/by-nc/2.0/legalcode.
### Table 1: Characteristics of included studies

| Study | Data collection period | Study design | Study population | Sample size (men/women) | Models | Outcomes |
|-------|-------------------------|--------------|------------------|-------------------------|--------|----------|
| Pandya et al\(^a\) | 2011 1988-94 | Retrospective | National health and nutrition examination survey III cohort | 5999 (3501/2498) | Framingham risk score, FRS (CVD)*, SCORE (low and high risk) | Cardiovascular disease mortality | 176 (118/58) |
| De la Iglesia et al\(^a\) | 2011 1995-2006 | Prospective | The Health Improvement Network cohort | 1 072 289 (529 506/542 783) | Framingham risk score, FRS (CVD)*, ASSIGN | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | 44 375 (26 202/18 173) |
| Barroso et al\(^b\) | 2010 No data | Retrospective | Cohort in Spain | 608 (263/345) | Framingham risk score, SCORE | Coronary heart disease (angina, fatal and non-fatal myocardial infarction), cardiovascular disease mortality | 57 (41/16) |
| Collins et al\(^a\) | 2010 1993-2008 | Prospective | The Health Improvement Network cohort | 1 583 106 (785 733/797 373) | Framingham risk score, QRISK1, QRISK2 | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | 71 465 (42 086/29 057) |
| Van der Heijden et al\(^m\) | 2009 1989-92 | Prospective | Cohort in Netherlands | 1125 (509/616) | Framingham risk score, SCORE | Coronary heart disease, cardiovascular disease mortality | 108 (coronary heart disease), 27 (fatal coronary heart disease) |
| Chen et al\(^m\) | 2009 2003-05 | Prospective | Cohort in Australia | 1998 (808/1190) | Framingham risk score, SCORE (low and high risk) | Cardiovascular disease mortality | 62 (36/26) |
| Collins et al\(^m\) | 2009 1995-2006 | Prospective | The Health Improvement Network cohort | 1 072 800 (529 813/542 987) | Framingham risk score, ASSIGN | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | 43 990 (25 963/18 027) |
| Woodward et al\(^m\) | 2009 1984-87, 1989, 1992, 1995 | Prospective | Scottish Heart Health Extended Cohort Study cohort | 13 060 (6509/6551) | Framingham risk score, ASSIGN | Cardiovascular disease mortality, coronary heart disease or cerebrovascular disease, CABG or PTCA | 2626 (1634/992) |
| Scheltens et al\(^m\) | 2008 1987-92 | Prospective | Cohort in Netherlands | 40 316 (18 814/21 502) | Framingham risk score, SCORE | Cardiovascular disease mortality | 256 (189/67) |
| Hippisley-Cox et al\(^x\) | 2008 1993-2008 | Prospective | QRESEARCH cohort | 752232 (374 469/375 763) | Framingham risk score, QRISK1, QRISK2 | Cardiovascular disease (coronary heart disease, stroke, transient ischaemic attack) | No data§ |
| Hippisley-Cox et al\(^x\) | 2007 1995-2007 | Prospective | QRESEARCH cohort | 614 553 (305 140/309 413) | Framingham risk score, QRISK1, ASSIGN | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | 30812 (17 705/13 107) |
| Mainous et al\(^a\) | 2007 1987-89 | Prospective | Atherosclerosis risk in communities study | 14 343 (6239/8104) | Framingham risk score, SCORE | Coronary heart disease (myocardial infarction, fatal coronary heart disease, cardiac procedure) | 1108 |
| Ridker et al\(^x\) | 2007 1992-2004 | Prospective | Women’s Health Study cohort | 8158 | Framingham risk score, Reynolds risk score | Cardiovascular disease mortality (myocardial infarction, ischemic stroke, coronary revascularisation, cardiovascular disease mortality) | 262 |
Table 1 (continued)

| Study            | Year | Data collection period          | Study design | Study population | Sample size (men/women) | Models                  | Outcomes                                           | No of events (men/women) |
|------------------|------|--------------------------------|--------------|------------------|-------------------------|-------------------------|---------------------------------------------------|--------------------------|
| Woodward et al²₁ | 2007 | 1984-87, 1998, 1992, 1995       | Prospective  | Cohort in Scotland | 1165 (743/422)          | Framingham risk score, ASSIGN | Cardiovascular disease mortality, coronary heart disease or cerebrovascular disease, CABG or PTCA | 1165 (743/422)            |
| Störk et al²₂     | 2006 | No data                        | Prospective  | Cohort in Netherlands | 403¶ | Framingham risk score, PROCAM | Cardiovascular disease and all cause mortality | 31¶                      |
| Cooper et al²₃    | 2005 | No data                        | Prospective  | Cohort in United Kingdom | 2732¶ | Framingham risk score, PROCAM | Coronary heart disease | 219¶                     |
| Ferrario et al²₀  | 2005 | 1982-96                        | Prospective  | Cuore study       | 6865¶ | Framingham risk score, PROCAM | Fatal and non-fatal major coronary heart disease | 312¶                     |
| Dunder et al²⁵    | 2004 | 1970-73                        | Prospective  | Uppsala Longitudinal Study of Adult Men cohort | 534¶ | Framingham risk score, PROCAM | Fatal and non-fatal myocardial infarction | 116¶                     |
| Empana et al²₈    | 2003 | 1991-93                        | Prospective  | Prospective Epidemiological Study of Myocardial Infarction cohorts: Northern Ireland; France | 2395¶; 7359¶ | Framingham risk score, PROCAM | Coronary heart disease (angina, fatal coronary heart disease, myocardial infarction) | 120¶; 197¶               |
| Assmann et al³¹   | 2002 | 1979-85                        | Prospective  | PROCAM cohort     | 5389¶ | Framingham risk score, PROCAM | Myocardial infarction or coronary heart disease mortality | 325¶                     |

SCORE=systematic coronary risk evaluation; ASSIGN=assessing cardiovascular risk to Scottish Intercollegiate Guidelines Network to assign preventative treatment score; QRISK1 and QRISK2=QRESEARCH cardiovascular risk algorithms; CABG=coronary artery bypass graft; PTCA=percutaneous transluminal coronary angioplasty; PROCAM=Prospective Cardiovascular Münster.

*Global Framingham risk score for total cardiovascular disease prediction.*
†Cohort subpopulation with normal glucose tolerance.
‡Derived from validation cohort.
§Data available from corresponding author.
¶Only males.
| Study                        | Year | Outcome                                                                 | Model                                      | AUC (95% CI)                                                                 | Overall |
|-----------------------------|------|--------------------------------------------------------------------------|--------------------------------------------|----------------------------------------------------------------------------|---------|
| Pandya et al<sup>25</sup>   | 2011 | Cardiovascular disease mortality                                         | Framingham risk score                       | 0.781 (0.738 to 0.823)                                                   | No data |
|                             |      |                                                                          | FRS (CVD)<sup>*</sup>                      | 0.776 (0.733 to 0.819)                                                   | No data |
|                             |      |                                                                          | SCORE                                      | Low risk: 0.785 (0.743 to 0.826), high risk: 0.785 (0.743 to 0.826)        | No data |
| De la Iglesia et al<sup>26</sup> | 2011 | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | Framingham risk score                       | 0.740 (0.736 to 0.744)<sup>†</sup>                                      | No data |
|                             |      |                                                                          | FRS (CVD)<sup>*</sup>                      | 0.752 (0.749 to 0.755)<sup>†</sup>                                      | No data |
|                             |      |                                                                          | ASSIGN                                     | 0.756 (0.753 to 0.759)<sup>†</sup>                                      | No data |
| Barroso et al<sup>27</sup>  | 2010 | Coronary heart disease (angina, fatal and non-fatal myocardial infarction), cardiovascular disease mortality | Framingham risk score                       | —                                                                        | 0.70 (0.63 to 0.78)                                           |
|                             |      |                                                                          | SCORE                                      | —                                                                        | 0.86 (0.77 to 0.96)                                           |
| Collins et al<sup>28</sup>  | 2010 | Cardiovascular disease (angina, myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | Framingham risk score                       | 0.75 (0.747 to 0.753)<sup>†</sup>                                       | No data |
|                             |      |                                                                          | QRSK1                                      | 0.771 (0.768 to 0.777)<sup>†</sup>                                      | No data |
|                             |      |                                                                          | QRSK2                                      | 0.773 (0.770 to 0.776)<sup>†</sup>                                      | No data |
| Van der Heijden et al<sup>29</sup> | 2009 | Coronary heart disease, coronary heart disease mortality                  | Framingham risk score                       | No data                                                                   | 0.68 (0.63 to 0.74)                                           |
|                             |      |                                                                          | SCORE                                      | No data                                                                   | 0.71 (0.66 to 0.76)                                           |
|                             |      |                                                                          | Framingham risk score                       | No data                                                                   | 0.71 (0.61 to 0.82)                                           |
|                             |      |                                                                          | SCORE                                      | No data                                                                   | 0.79 (0.70 to 0.87)                                           |
| Chen et al<sup>30</sup>     | 2009 | Cardiovascular disease mortality                                         | Framingham risk score                       | 0.72 (0.65 to 0.80)                                                      | No data |
|                             |      |                                                                          | SCORE                                      | Low risk: 0.75 (0.68 to 0.83), high risk: 0.75 (0.68 to 0.82)             | No data |
| Collins et al<sup>31</sup>  | 2009 | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | Framingham risk score                       | 0.752 (0.749 to 0.755)<sup>†</sup>                                       | No data |
|                             |      |                                                                          | QRSK1                                      | 0.770 (0.766 to 0.774)<sup>†</sup>                                       | No data |
|                             |      |                                                                          | QRSK2                                      | 0.789 (0.785 to 0.793)<sup>†</sup>                                       | No data |
| Woodward et al<sup>32</sup> | 2009 | Cardiovascular disease mortality, coronary heart disease or cerebrovascular disease, CAGB or PTCA | Framingham risk score                       | 0.7183 (0.7154 to 0.7213)                                               | No data |
|                             |      |                                                                          | ASSIGN                                     | 0.7248 (0.7216 to 0.7279)                                               | No data |
| Scheltens et al<sup>33</sup> | 2008 | Cardiovascular disease mortality                                         | Framingham risk score                       | No data                                                                   | 0.86 (0.84 to 0.88)                                           |
|                             |      |                                                                          | SCORE                                      | No data                                                                   | 0.85 (0.83 to 0.87)                                           |
| Hippisley-Cox et al<sup>34</sup> | 2008 | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | Framingham risk score                       | 0.779 (0.776 to 0.782)                                                  | No data |
|                             |      |                                                                          | QRSK1                                      | 0.788 (0.786 to 0.791)                                                  | No data |
|                             |      |                                                                          | QRSK2                                      | 0.792 (0.789 to 0.794)                                                  | No data |
| Hippisley-Cox et al<sup>35</sup> | 2007 | Cardiovascular disease (myocardial infarction, coronary heart disease, stroke, transient ischaemic attack) | Framingham risk score                       | 0.7598 (0.756 to 0.764)<sup>†</sup>                                      | No data |
|                             |      |                                                                          | QRSK1                                      | 0.7674 (0.763 to 0.772)<sup>†</sup>                                      | No data |
|                             |      |                                                                          | QRSK2                                      | 0.7879 (0.785 to 0.791)<sup>†</sup>                                      | No data |
|                             |      |                                                                          | ASSIGN                                     | 0.7644 (0.760 to 0.769)<sup>†</sup>                                      | No data |
| Mainous et al<sup>36</sup>  | 2007 | Coronary heart disease (myocardial infarction, fatal coronary heart disease, cardiac procedure) | Framingham risk score                       | 0.691 (0.670 to 0.712)                                                  | No data |
|                             |      |                                                                          | SCORE                                      | 0.619 (0.597 to 0.641)                                                  | No data |
Table 2 (continued)

| Study                  | Year | Outcome                                                                 | Model                          | AUC (95% CI)         | Men                          | Women                        | Overall                      |
|------------------------|------|--------------------------------------------------------------------------|-------------------------------|----------------------|------------------------------|------------------------------|------------------------------|
| Ridker et al 2007      |      | Cardiovascular disease (myocardial infarction, ischaemic stroke, coronary revascularisation, cardiovascular disease mortality) | Framingham risk score         | NA                   | 0.787 (0.754 to 0.820)† NA   |                              |                              |
|                        |      |                                                                          | Reynolds risk score           | NA                   | 0.808 (0.776 to 0.840)† NA   |                              |                              |
| Woodward et al 2007    |      | Cardiovascular disease mortality, coronary heart disease or cerebrovascular disease, CABG or PTCA | Framingham risk score         | 0.716 (0.694 to 0.738)† | 0.741 (0.720 to 0.762)† No data |                              |                              |
|                        |      |                                                                          | ASSIGN                        | 0.727 (0.706 to 0.748)† | 0.765 (0.744 to 0.786)† No data |                              |                              |
| Störk et al 2006       |      | Cardiovascular disease and all cause mortality                            | Framingham risk score         | 0.60 (0.49 to 0.69)   | NA                           | NA                           |                              |
|                        |      |                                                                          | PROCAM                        | 0.55 (0.45 to 0.65)   | NA                           | NA                           |                              |
| Cooper et al 2005      |      | Coronary heart disease                                                  | Framingham risk score         | 0.62 (0.58 to 0.66)   | NA                           | NA                           |                              |
|                        |      |                                                                          | PROCAM                        | 0.63 (0.59 to 0.67)   | NA                           | NA                           |                              |
| Ferrario et al 2005    |      | Fatal and non-fatal major coronary heart disease                         | Framingham risk score         | 0.723 (0.670 to 0.779) | NA                           | NA                           |                              |
|                        |      |                                                                          | PROCAM                        | 0.735 (0.678 to 0.790) | NA                           | NA                           |                              |
| Dunder et al 2004      |      | Fatal and non-fatal myocardial infarction                               | Framingham risk score         | 0.61 (0.55 to 0.67)†  | NA                           | NA                           |                              |
|                        |      |                                                                          | PROCAM                        | 0.63 (0.57 to 0.69)†  | NA                           | NA                           |                              |
| Empana et al 2003      |      | Coronary heart disease (angina, fatal coronary heart disease, myocardial infarction) | Framingham risk score‡         | 0.66 (0.606 to 0.714)† | NA                           | NA                           |                              |
|                        |      |                                                                          | PROCAM‡                       | 0.61 (0.555 to 0.665)† | NA                           | NA                           |                              |
|                        |      |                                                                          | Framingham risk score§         | 0.68 (0.638 to 0.722)† | NA                           | NA                           |                              |
|                        |      |                                                                          | PROCAM§                       | 0.64 (0.598 to 0.682)† | NA                           | NA                           |                              |
| Assmann et al 2002     |      | Myocardial infarction or cardiovascular disease mortality                | Framingham risk score         | 0.778 (0.748 to 0.808)† | NA                           | NA                           |                              |
|                        |      |                                                                          | PROCAM                        | 0.824 (0.796 to 0.852)† | NA                           | NA                           |                              |

SCORE—systematic coronary risk evaluation; ASSIGN—assessing cardiovascular risk to Scottish Intercollegiate Guidelines Network to assign preventative treatment score; CABG—coronary artery bypass graft; PTCA—percutaneous transluminal coronary angioplasty; NA—not applicable; PROCAM—Prospective Cardiovascular Münster score.

*Global Framingham risk score for total cardiovascular disease prediction.
†Confidence intervals calculated as described in Hanley and McNeil.‡‡
§Northern Ireland cohort.
§§France cohort.
| Study                     | Model          | Comparator | Performed better than comparator(s)* | Involved some of same authors | Involved independent authors |
|--------------------------|----------------|------------|--------------------------------------|-------------------------------|------------------------------|
| Hippisley-Cox et al      | QRISK2         | Framingham risk score, QRISK1 | Yes                                   | None                          | QRISK2>Framingham risk score |
|                          |                |            |                                      |                               | QRISK1>Framingham risk score |
|                          | QRISK1         | Framingham risk score, ASSIGN    | Yes                                   | QRISK1>Framingham risk score  |
|                          |                |            |                                      |                               | QRISK1>Framingham risk score  |
| Ridker et al             | Reynolds risk score | Framingham risk score | Yes                                   | None                          | None                         |
| Woodward et al           | ASSIGN         | Framingham risk score | Yes                                   | ASSIGN>Framingham risk score  |
|                          |                |            |                                      |                               | ASSIGN>Framingham risk score  |
| Assmann et al            | PROCAM         | Framingham risk score | Yes                                   | None                          | PROCAM>Framingham risk score |
|                          |                |            |                                      |                               | PROCAM>Framingham risk score |

QRISK1 and QRISK2=QRESEARCH cardiovascular risk algorithms; ASSIGN=assessing cardiovascular risk to Scottish Intercollegiate Guidelines Network to assign preventative treatment score; PROCAM=Prospective Cardiovascular Münster score.

*Better performance of models is based on point estimates for area under the receiver operating characteristic curve.

**Table 3** Potential optimism bias

---

BMJ: first published as 10.1136/bmj.e3318 on 24 May 2012. Downloaded from http://www.bmj.com on 29 October 2023 by guest. Protected by copyright.