2019

Equalization of Four Cardiovascular Risk Algorithms After Systematic Recalibration: Individual-Participant Meta-Analysis of 86 Prospective Studies

Lisa Pennells
University of Cambridge, Cambridge, United Kingdom

Stephen Kaptoge
University of Cambridge, Cambridge, United Kingdom

Angela Wood
University of Cambridge, Cambridge, United Kingdom

Mike Sweeting
University of Cambridge, Cambridge, United Kingdom

Xinghui Zhao
University of Cambridge, Cambridge, United Kingdom

See next page for additional authors

Let us know how access to this document benefits you.

Follow this and additional works at: https://pdxscholar.library.pdx.edu/commhealth_fac

Part of the Community Health Commons

Citation Details

This Article is brought to you for free and open access. It has been accepted for inclusion in Community Health Faculty Publications and Presentations by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
Authors
Lisa Pennells, Stephen Kaptoge, Angela Wood, Mike Sweeting, Xinghui Zhao, Carlos J. Crespo, and +

This article is available at PDXScholar: https://pdxscholar.library.pdx.edu/commhealth_fac/94
Equalization of four cardiovascular risk algorithms after systematic recalibration: individual-participant meta-analysis of 86 prospective studies


* Corresponding author. Tel: 01223 748659, Fax: 01223 748658, Email: ed303@medschl.cam.ac.uk
† These authors contributed equally to this article.
‡ Investigators of the Emerging Risk Factors Collaboration are listed at the end of this manuscript.
© The Author(s) 2018. Published by Oxford University Press on behalf of the European Society of Cardiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Aims  
There is debate about the optimum algorithm for cardiovascular disease (CVD) risk estimation. We conducted head-to-head comparisons of four algorithms recommended by primary prevention guidelines, before and after ‘recalibration’, a method that adapts risk algorithms to take account of differences in the risk characteristics of the populations being studied.

Methods and results  
Using individual-participant data on 360,737 participants without CVD at baseline in 86 prospective studies from 22 countries, we compared the Framingham risk score (FRS), Systematic COronary Risk Evaluation (SCORE), pooled cohort equations (PCE), and Reynolds risk score (RRS). We calculated measures of risk discrimination and calibration, and modelled clinical implications of initiating statin therapy in people judged to be at ‘high’ 10 year CVD risk. Original risk algorithms were recalibrated using the risk factor profile and CVD incidence of target populations. The four algorithms had similar risk discrimination. Before recalibration, FRS, SCORE, and PCE overpredicted CVD risk on average by 10%, 52%, and 41%, respectively, whereas RRS under-predicted by 10%. Original versions of algorithms classified 29–39% of individuals aged ≥40 years as high risk. By contrast, recalibration reduced this proportion to 22–24% for every algorithm. We estimated that to prevent one CVD event, it would be necessary to initiate statin therapy in 44–51 such individuals using original algorithms, in contrast to 37–39 individuals with recalibrated algorithms.

Conclusion  
Before recalibration, the clinical performance of four widely used CVD risk algorithms varied substantially. By contrast, simple recalibration nearly equalized their performance and improved modelled targeting of preventive action to clinical need.

Keywords  
Cardiovascular disease • Risk prediction • Risk algorithms • Calibration • Discrimination

Introduction  
A key strategy in the primary prevention of cardiovascular disease (CVD) is the use of risk prediction algorithms to target preventive interventions on people who should benefit from them most. There is, however, debate about the optimum algorithm for CVD risk estimation. The 2013 guidelines of the American College of Cardiology/American Heart Association (ACC/AHA) have recommended the Framingham risk score (FRS), Systematic COronary Risk Evaluation (SCORE), pooled cohort equations (PCE), and Reynolds risk score (RRS). We conducted head-to-head comparisons of four algorithms recommended by primary prevention guidelines, before and after ‘recalibration’, a method that adapts risk algorithms to take account of differences in the risk characteristics of the populations being studied. Few studies have compared recalibrated versions of algorithms systematically across many populations.

Our study, therefore, aimed to address two sets of questions. First, how do risk prediction algorithms differ in terms of predictive accuracy and clinical performance when evaluated in the same population? We chose algorithms that have been recommended by a guideline statement and could be evaluated with the information available in our consortium dataset. Hence, we conducted head-to-head comparisons of original versions of four risk algorithms (FRS, SCORE, PCE, and RRS), evaluating them using measures of predictive accuracy (e.g. discrimination, calibration) as well as clinical performance (e.g. modelled the potential impact of initiating statin therapy as recommended by primary prevention CVD guidelines). The second set of questions is: what is the clinical impact of adjusting these algorithms to local and contemporary circumstances, and how do they then compare to each other? To address them, we recalibrated these algorithms using CVD event rates and risk factor values of the target populations, and compared the performance of the original and recalibrated versions of algorithms across multiple settings.

Methods  
Data sources  
We analysed data from the Emerging Risk Factors Collaboration (ERFC), a consortium of prospective cohort studies with information on a variety of risk factors. Prospective cohort studies were included in this analysis if they met all the following criteria: (i) had not contributed data to the development of any of the risk prediction algorithms studied in this analysis; (ii) had recorded information on risk factors necessary to calculate algorithms (i.e. age, sex, smoking status, history of diabetes, systolic blood pressure, total and high-density lipoprotein cholesterol,
The statistical analysis included participants aged between 40 and 79 years, excluding those with a known history of CVD at baseline (i.e., coronary heart disease (CHD), other heart disease, stroke, transient ischaemic attack, peripheral vascular disease, atrial fibrillation, heart failure, or any cardiovascular surgery), as defined by each study. For each participant, we used original versions of FRS, SCORE, PCE, and RRS to calculate the predicted 10-year risk of CVD outcomes (Supplementary material online, Table S1). To enable comparison with the three other risk prediction algorithms evaluated in this study, we used a rescaled version of the FRS algorithm which predicts non-fatal MI, fatal CHD, or any stroke (rather than the broader CVD outcome it was originally derived for). For SCORE, we used relevant high or low-risk versions depending on the geographical location of the cohort as recommended by the ESC guidelines. Analyses involving RRS were performed in a subset of participants who had information available on C-reactive protein, family history of premature MI, and HbA1c (if female and with diabetes) (Supplementary material online, Table S1).

To help provide systematic evaluation of the four risk algorithms to predict relevant CVD endpoints, we used the following outcome definitions. The principal outcome was the composite of CVD events during the initial 10 year period of follow-up as defined by each algorithm (the algorithm-specific outcome): first onset of non-fatal MI, fatal CHD, or any stroke for FRS and PCE; non-fatal MI, fatal CHD or any stroke, coronary revascularization, or any CVD death for RRS; fatal CVD for SCORE (Supplementary material online, Table S3). The secondary outcome was a ‘common’ CVD outcome, defined as the composite of non-fatal MI, fatal CHD, or any stroke, adopting the definition of the 2013 ACC/AHA guidelines (and used by PCE and FRS). Outcomes were censored if a participant was lost to follow-up, died from non-CVD causes, or reached 10 years of follow-up. Participants contributed only the first non-fatal or fatal CVD event (i.e., deaths preceded by non-fatal CVD events were not included) except in the case of the SCORE-specific outcome, for which all fatal CVD events were included.

We assessed risk discrimination using the C-index which estimates the probability of correctly predicting who will have a CVD event first in a randomly selected pair of participants. The C-index calculation was stratified by sex and involved a two-stage approach, with estimates calculated separately within each study before pooling across studies weighting by the number of contributing events. We assessed calibration of risk algorithms for each algorithm-specific outcome by comparing predicted and observed risks calculated for groups of participants defined by 5-year age categories and calculating goodness of fit tests. Supplementary material online, Table S3 provides further details of the methods used to assess calibration. We recalibrated each algorithm as shown in Supplementary material online, Figure S1 and Supplementary material online, Appendix S2. Our approach involved adaptation of original risk algorithms using the risk factor profile and CVD incidence of target populations. Recalibration to CVD incidence involved two approaches. First, we recalibrated each algorithm to predict incidence of the endpoint it was derived to predict (the algorithm-specific outcome). Second, to enable head-to-head comparisons, we recalibrated SCORE and RRS to the common CVD outcome used by FRS and PCE, as mentioned above. Only studies with at least 10 years of follow-up were used in analyses involving recalibration, or assessment of calibration.

To assess the clinical implications of using different algorithms to initiate statin therapy in those whose 10-year CVD risk exceeds a given threshold (as recommended by several CVD primary prevention guideline statements), we estimated the number of individuals who would be eligible for treatment and the potential cases avoided. First, we assumed CVD risk assessment for a population of 100,000 men and women aged 40 years without CVD at baseline and not already taking statins or meeting guideline recommendations for statin treatment (i.e., people without a history of diabetes or CVD and with low-density lipoprotein (LDL) cholesterol <190 mg/dL). Second, we assumed the same age structure of a standard population of the United States. Third, we assumed age- and sex-specific incidence rates for CVD events as in the current study. Fourth, we assumed statin allocation according to the threshold of predicted 10-year CVD risk recommended by 2013 ACC/AHA guidelines for first-onset fatal and non-fatal CVD events (i.e., >7.5%), or by the 2016 ESC Guidelines for fatal CVD (i.e., >5%). Fifth, we assumed CVD risk reductions of 20% with statin treatment in people without a history of CVD, as reported by the Cholesterol Treatment Trials’ Collaboration. We also compared categorization of participants across different algorithms before and after their recalibration using the net reclassification improvement (NRI).

Analyses were performed using Stata version 14. P-values are two-sided. The study was designed and conducted by this collaboration’s academic coordinating centre, and was approved by the Cambridgeshire Ethics Review Committee. The funders had no scientific role in the study.

**Results**

We analysed data on 360,737 participants without prior CVD who were recruited into 86 prospective cohorts between the years 1963 and 2003 (Supplementary material online, Table S2). The mean (standard deviation) age at baseline was 59 (8) years; 53% were male. Sixty-nine percent of the participants were recruited in European countries, 18% in North America, and the remainder mostly in Japan and Australia. Median (5th–95th percentile) follow-up was 10.2 (3.4–21.3) years, and during the initial 10 years of follow-up (3.1 million person-years at risk), 14,564 incident CVD events were recorded according to our common and FRS/PCE CVD definition, including 9,259 CHD events and 5,305 stroke events. At baseline, the median (5th–95th percentile) predicted 10-year CVD risks were 5.54% (1.02–23.34%) using FRS, 2.49% (0.13–23.25%) using SCORE, and 6.43% (0.69–33.33%) using PCE (Table 1). Baseline characteristics for the subset of participants with information on the RRS are presented in Supplementary material online, Table S4.
Table 1  Baseline characteristics and predicted 10 year cardiovascular disease risk relevant to assessed algorithms

<table>
<thead>
<tr>
<th>Baseline characteristic</th>
<th>Mean (SD) or n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at survey (years)</td>
<td>59 (8.0)</td>
</tr>
<tr>
<td>Males</td>
<td>189 342 (52.5%)</td>
</tr>
<tr>
<td>Current smoking</td>
<td>98 593 (27.3%)</td>
</tr>
<tr>
<td>History of diabetes</td>
<td>16 758 (4.6%)</td>
</tr>
<tr>
<td>Systolic blood pressure (mmHg)</td>
<td>132 (19)</td>
</tr>
<tr>
<td>Total cholesterol (mmol/L)</td>
<td>5.83 (1.08)</td>
</tr>
<tr>
<td>HDL cholesterol (mmol/L)</td>
<td>1.33 (0.38)</td>
</tr>
<tr>
<td>Total/HDL cholesterol ratio</td>
<td>4.50 (1.61)</td>
</tr>
<tr>
<td>Hypertension medication</td>
<td>37 960 (10.5)</td>
</tr>
<tr>
<td>Lipid lowering medication</td>
<td>7929 (5.1)</td>
</tr>
<tr>
<td>Predicted 10 year risk (%)</td>
<td>median (5th-95th percentiles)</td>
</tr>
<tr>
<td>Framingham risk score (FRS)</td>
<td>5.54% (1.02–23.34)</td>
</tr>
<tr>
<td>Systematic COronary Risk (SCORE)</td>
<td>2.49% (0.13–23.25)</td>
</tr>
<tr>
<td>Evaluation (SCORE)</td>
<td></td>
</tr>
<tr>
<td>Pooled cohort equations (PCE)</td>
<td>6.43% (0.69–33.33)</td>
</tr>
</tbody>
</table>

Predicted 10 year risk: FRS, SCORE, PCE, and RRS did not differ. However, the extent and direction of mis-calibration varied substantially across individual cohorts, ranging from more than 50% underestimation to >400% overestimation of risk (Supplementary material online, Figures S7 and S8). Heterogeneity in calibration could not be systematically explained by broad geographical region but was partially explained by year of baseline screening (Supplementary material online, Figure S9).

After recalibration of algorithms to the incidence of the common CVD outcome and risk factor distribution of the cohorts contributing to the current analysis, the distribution of predicted 10 year CVD risk was similar across the four algorithms we studied (Supplementary material online, Figure S10), yielding good calibration for each algorithm (Supplementary material online, Figure S11). Risk discrimination did not change with recalibration since ranking of participant risk is unaffected by the recalibration methods used (Supplementary material online, Figure S1 and Appendix S3).

Discrimination and calibration

When using algorithm-specific CVD outcomes, each algorithm provided broadly similar discrimination, with absolute C-index values ranging from 0.7010 to 0.7605. The pooled cohort equations provided somewhat greater risk discrimination than FRS or SCORE for all algorithm-specific outcomes, with differences in overall C-index compared with FRS between 0.0039 and 0.0131 (P < 0.001 when testing the null hypothesis of no difference between C-indices; Figure 1). Differences were greater for women than men, but similar among participants from European and North American cohorts (Supplementary material online, Figure S2). A similar pattern was observed in analyses restricted to participants with complete data enabling calculation of RRS (Supplementary material online, Figure S3). Differences in the C-index among algorithms were not affected by study recruitment periods (Supplementary material online, Figure S4).

For each algorithm-specific outcome, on average across cohorts the predicted 10 year risk was 1.10 times observed risk for FRS, 1.52 for SCORE, 1.41 for PCE, and 0.90 for RRS (P < 0.0001 for goodness of fit/calibration for all algorithms; Figure 2 and Supplementary material online, Figures S5 and S6). On average the extent of relative mis-calibration was similar in men and women, and across all ages for SCORE and PCE (Supplementary material online, Figure S5) which translated to greater discrepancy between absolute predicted and observed risks at older ages when using these algorithms (Supplementary material online, Figure S6). Framingham risk score tended to over-predict in men and younger women but to under-predict in older women. Reynolds risk score underestimated risk somewhat in men, but on average was well calibrated in women (Figure 2, Supplementary material online, Figures S5 and S6). The extent and direction of mis-calibration varied substantially across individual cohorts, ranging from more than 50% underestimation to >400% overestimation of risk (Supplementary material online, Figures S7 and S8). Heterogeneity in calibration could not be systematically explained by broad geographical region but was partially explained by year of baseline screening (Supplementary material online, Figure S9).

Estimates of clinical performance

We initially conducted modelling that employed original versions of the four CVD risk algorithms we studied; was weighted to represent the age and sex distribution of a standard US population ≥40 years; focused on individuals not already taking or eligible for statin treatment (i.e. people without a history of diabetes or CVD and with LDL <190 mg/dL); and defined the threshold for initiation of statin treatment as an absolute 10 year risk of >7.5% for FRS, PCE, and RRS, and ≥5% for SCORE (‘high risk’).

Under this scenario, we estimated that the proportion of individuals classified as high-risk (i.e. eligible for statin treatment) was 32% with FRS, 29% with SCORE, 39% with PCE, and 32% with RRS (Supplementary material online, Table S5 and Figure 3). By contrast, after recalibration (using algorithmic-specific CVD endpoints), FRS, SCORE, PCE, and RRS predicted CVD outcomes more accurately, classified lower proportions of people as high risk, and identified higher proportions of CVD events among people classified as high risk. After further recalibration to the common CVD endpoint, the proportion of individuals classified as high risk lowered to a near uniform level (22%, 22%, 24%, and 23% with FRS, SCORE, PCE, and RRS, respectively). Of those classified as high risk by the original versions of algorithms, 11% later developed a first CVD event within 10 years (i.e. the positive predictive value was 11%, 11%, 10%, and 11%, respectively). By contrast, it was 13% with the recalibrated algorithms (Supplementary material online, Table S5).

Based on these estimates, we calculated that to prevent one CVD event when using original versions of FRS, SCORE, PCE, or RRS it would be necessary to initiate statin therapy in 46, 44, 51, or 45 individuals, respectively (following screening of 145, 150, 131, or 142 individuals, respectively; Figure 3 and Supplementary material online, Table S5). By contrast, when using any of the recalibrated algorithms, one CVD event could be prevented by initiating statin therapy in 38 participants (following screening of 174, 171, 160, or 165 individuals, respectively). Similar findings to those observed above were noted in analyses that used a range of treatment thresholds different from those in current guidelines (Figure 3) with the divergent clinical performance of original algorithms converging to become almost identical at any treatment threshold after recalibration to a common CVD endpoint.
We then modelled the concordance of statin treatment decisions based on use of these algorithms. Before recalibration, 41% of all individuals were at high risk with at least one of the four algorithms and 58% of these (24% of all individuals) were at high risk with all four. By contrast, after recalibration to our common CVD outcome, 28% of individuals were at high risk with at least one algorithm and 63% of these (18% of all individuals) were at high risk with all four (Supplementary material online, Figure S12). Discordance in treatment decisions before recalibration tended to be greatest when comparing SCORE to the other algorithms (Figure 4). For example, in pairwise comparisons between FRS and SCORE, in every 100 000 people screened 36 794 would be classified as high risk with either FRS or SCORE and 24 157 (66% of these) would be classified as high risk with both FRS and SCORE. By contrast, after recalibration, 18 716 (76%) of the 24 708 individuals at high risk with either FRS or SCORE would be at high risk with both algorithms (Figure 4). This greater concordance between algorithms in identifying those at high risk was also illustrated by a decrease in the NRI among both cases and event-free participants after recalibration (Supplementary material online, Table S6) and greater agreement between the absolute risk predictions (Supplementary material online, Figure S13).

**Discussion**

In an analysis of individual-participant data on over 350 000 people without a history of CVD at baseline, we systematically evaluated several risk algorithms recommended by North American and European guidelines for primary prevention of CVD. Our study’s main finding was that the clinical performance of four widely used risk algorithms varied substantially, predominantly due to differing extent of calibration. By contrast, we observed only slight differences among the algorithms in relation to risk discrimination (a measure of predictive accuracy that is not influenced by the extent of model calibration). After recalibration, however, the performance of the four algorithms was essentially equalized. Our modelling suggested, therefore, that targeting of CVD preventive action to clinical need would improve considerably due to higher accuracy of individual risk predictions. A key implication of these results is that CVD primary prevention guidelines should shift away from debates about the relative merits of particular risk algorithms and, instead, achieve consensus about the need for more widespread use of any recalibrated algorithm.

Our findings have suggested that effective recalibration can be achieved through the use of simple methods that can be applied using
aggregate level data on CVD event rates and average risk factor values for a target population to be screened. To scale this approach for clinical and public health purposes, cardiovascular bodies might facilitate the collation and regular updating of national and regional age- and sex-specific CVD event rates and risk factor data, including for particular geographical areas and ethnic groups with distinctive CVD event rates and risk factors values. This information could then be embedded in user-friendly risk prediction tools (e.g. online risk calculators or electronic health records systems), enabling regular and simple recalibration, as previously described.28,29 An alternative approach is the periodic development of new risk algorithms, although it would be more costly and time-consuming than recalibration because it entails launch of large new cohort studies and their long-term follow-up.

In contrast with previous analyses of simulated data, studies in single populations, or comparisons of risk scores without recalibration,17–20,30–35 our study directly compared original and recalibrated versions of four algorithms used across many different populations, providing the first demonstration of the extent of CVD risk prediction improvement achievable through recalibration. For example, following recalibration we observed that the proportion of individuals classified as high risk reduced from about 40% to 23%, and the number of individuals needed to initiate statin therapy to prevent one event reduced from between 44–51 to around 38. However, our modelling reflects the average improvement that could be achieved in countries or regions where mis-calibration is more extreme could potentially be much greater.

Our approach to recalibration was distinctive in two ways. First, it extended previous recalibration methods36 by using age groups instead of categories of predicted risk, which allows direct application to population data that are routinely recorded. Second, it differed from other recalibration methods proposed for specific CVD risk algorithms28,29 by providing a simpler procedure applicable to algorithms derived using any type of statistical model. Because we studied participant-level data from cohorts with prolonged follow-up, we could adopt a uniform approach to statistical analyses and conduct time-to-event analyses. To avoid providing over-optimistic assessment of algorithm performance, we omitted cohorts that had previously contributed data to the derivation of the risk algorithms we studied. Our clinical modelling was robust to different scenarios. The generalizability of our findings was enhanced by inclusion of several dozen population cohorts in 22 countries, mostly in Europe and North America, and the broad range in baseline year of recruitment across studies.

Our study had potential limitations. Because we used data from the target cohorts themselves to recalibrate algorithms, the benefits of recalibration could have been exaggerated (albeit in a manner that would have affected each algorithm identically). Conversely, inaccuracy in CVD ascertainment in contributing cohorts would tend to worsen the apparent performance of algorithms (again, affecting each algorithm identically).37 Our modelling could have over-estimated potential benefits of statin therapy because not all people eligible for statins will receive them or be willing or able to take them. On the
other hand, greater clinical impact than suggested by our modelling would be estimated if we had used less conservative assumptions (e.g. use of more efficacious statin regimens or additional treatments; longer time horizons; and lifestyle changes). We did not formally incorporate the impact of the potential hazards of statins into our modelling. We had incomplete information on medication use (such as statins and antihypertensive drugs) or cardiovascular intervention (such as coronary revascularization) during follow-up, which may have influenced our estimates of the observed CVD risk. Revascularization endpoints may have been differentially recorded.

Figure 3 Estimated public health impact with screening using original and recalibrated cardiovascular disease risk prediction algorithms over a range of risk thresholds in a standard US population of 100,000 people aged over 40 years. Cardiovascular disease includes fatal coronary heart disease, fatal, and non-fatal myocardial infarction and any stroke. FRS, Framingham risk score; PCE, pooled cohort equations; RRS, Reynolds risk score; SCORE, Systematic COronary Risk Evaluation.
across studies, which may have impacted on our assessment of calibration of the original RRS. There is, as yet, no randomized evidence that CVD risk assessment translates into CVD prevention.38

**Conclusion**

Whereas the performance of the original versions of four widely used CVD risk algorithms varied substantially, simple recalibration essentially equalized them and improved targeting of CVD preventive action to clinical need. This study supports the concept of using regularly recalibrated risk algorithms in routine clinical practice.

**Investigators of the Emerging Risk Factors Collaboration**

Data Management Team
Thomas Bolton, Sarah Spackman, and Matthew Walker.

Co-ordinating Centre

Supplementary material
Supplementary material is available at European Heart Journal online.

Funding
The work of the co-ordinating centre was funded by the UK Medical Research Council (G0800270), British Heart Foundation (SP/09/
from Novartis, Kowa, Pfizer, NHLBI, outside the submitted work; he is listed as a co-inventor on patents held by the Brigham and Women’s Hospital that relate to the use of inflammatory biomarkers in cardiovascular disease and diabetes that have been licensed to AstraZeneca and Seimens; R.R. reports grants, personal fees and non-financial support from Sanofi, MSD, Agen, Physigenex, AstraZeneca, Novo Nordisk, Janssen, Eli Lilly, Abbott, Medtronic, Servier, outside the submitted work; V.S. reports personal fees from Novo Nordisk outside the submitted work; N.S. reports grants and personal fees from Agen, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Novo Nordisk, Sanofi, outside the submitted work; S.G.T. reports grants from UK Medical Research Council, and British Heart Foundation, during the conduct of the study; P.W. reports personal fees from Novartis Pharmaceuticals, outside the submitted work; M.W. reports personal fees from Agen, outside the submitted work. The other authors declare no competing interests.

References


