Performance of the ASSIGN cardiovascular disease risk score on a UK cohort of patients from general practice

Beatriz de la Iglesia,[1] John F Potter,[2] Neil R Poulter,[3] Margaret M Robins,[1] Jane Skinner[2]

See Editorial, p 442 and Featured correspondence, p 515

[1]School of Computing Sciences, University of East Anglia, Norwich, UK
[2]School of Medicine, Health Policy and Practice, University of East Anglia, Norwich, UK
[3]ICCH, Imperial College London, London, UK

Correspondence to
Beatriz de la Iglesia, School of Computing Sciences, University of East Anglia, UEA Campus, Norwich, NR4 7TJ, UK; bli@cmp.uea.ac.uk

Accepted 28 September 2010
Published Online First 20 November 2010

ABSTRACT

Objective To evaluate the performance of ASSIGN against the Framingham equations for predicting 10 year risk of cardiovascular disease in a UK cohort of patients from general practice and to make the evaluation comparable to an independent evaluation of QRISK on the same cohort.

Design Prospective open cohort study.

Setting 288 practices from England and Wales contributing to The Health Improvement Network (THIN) database.

Participants Patients registered with 288 UK practices for some period between January 1995 and March 2006. The number of records available was 1 787 169.

Main outcome measures First diagnosis of myocardial infarction, coronary heart disease, stroke and transient ischaemic attacks recorded.

Methods We implemented the Anderson Framingham Coronary Heart Disease and Stroke models, ASSIGN, and a more recent Framingham Cox proportional-hazards model and analysed their calibration and discrimination.

Results Calibration showed that all models tested overestimated risk particularly for men. ASSIGN showed better discrimination with higher AUROC (0.756/0.792 for men/women), D statistic (1.35/1.58 for men/women), and R^2 (30.47%/37.39% for men/women). The performance of ASSIGN was comparable to that of QRISK on the same cohort. Models agreed on 93—97% of categorical (high/lower) risk assessments and when they disagreed, ASSIGN was often closer to the estimated Kaplan-Meier incidence. ASSIGN also provided a steeper gradient of deprivation and discriminated between those with and without recorded family history of CVD. The estimated incidence was twice/three times as high for women/men with a recorded family history of CVD.

Conclusions For systematic CVD risk assessment all models could usefully be applied, but ASSIGN improved on the gradient of deprivation and accounted for recorded family history whereas the Framingham equations did not. However, all models display relatively low specificity and sensitivity. An additional conclusion is that the recording of family history of CVD in primary care databases needs to improve given its importance in risk assessment.

INTRODUCTION

Current clinical guidelines for cardiovascular disease (CVD) such as those of the Joint British Societies[1] propose that CVD prevention should focus on all those at high risk: (i) people with established CVD; (ii) people with diabetes, and (iii) apparently healthy individuals at high estimated risk of CVD (CVD risk of ≥20% over 10 years). The latter group requires an accurate method for calculating CVD risk.

The Joint British Societies' guidelines published in 2005 recommended the use of the Framingham 1991 10-year risk equations (referred to in this paper as Anderson Framingham) to assess CVD risk.[2 3] The NICE guidelines[4] initially recommended Framingham but were updated in 2010 to recommend that Framingham should be considered as one of the possible equations to use. In this context, it is important for healthcare professionals to be aware of the different equations available and their performance so they can make an informed choice.

The Framingham equations are widely used throughout the world and have been tested in many different situations and adapted ('calibrated') accordingly. However, their application is not without problems. A systematic review by Brindle *et al*[5] has shown that the accuracy of the Framingham estimates cannot be assumed and that it relates to the background risk of the population to which they are being applied. There have also been calls for additional factors to be included into the calculation, notably ethnicity,[6] family history[7 8] and socio-economic factors.[9 10] Development of models from the Framingham Cohort continues to date and a new Cox proportional-hazards model for prediction of a first CVD event has been published.[11] We will refer to this as the Cox Framingham model.

Recently there have been two major UK contributions to the development of accurate CVD risk scores: the ASSIGN algorithm,[12] a model derived from the Scottish Heart Health Extended Cohort (SHHEC)[13] and the QRISK algorithm.[14 15] QRISK is a significant development because it was derived using a large UK-based primary care database similar to THIN instead of a cohort study. Primary care datasets contain large amounts of data and are increasingly finding a use in research but contain many missing values and other data quality problems. In cohort studies, such as the SHHEC or the Framingham cohort, the data are collected for particular research purposes and can therefore be of higher quality. Unlike the Framingham models, both ASSIGN and QRISK include measures of social deprivation and family history.

The NICE guidelines[4] identified 'an urgent need to establish which score is most acceptable for use in the population of England and Wales' and called

for research 'to assess the use of ASSIGN in UK populations outside Scotland'. An independent validation of QRISK[16] was published recently which compared the performance of QRISK against the Anderson and Cox Framingham equations. The comparison did not include the ASSIGN risk equation. The aim of this paper is to provide an independent and comparable validation of ASSIGN against the Framingham equations using the THIN dataset which contains patients from England and Wales. We are unable to produce a direct comparison with the QRISK results as we are not able to obtain them because the QRISK algorithm remains unpublished (ie, mean values for clinical characteristics used in the algorithm and details of the fractional polynomial models have not been published). We therefore present an indirect comparison by using similar analysis methods and presentation style.

METHODS

Data source

The data used in this study were extracted from the THIN database by the Information Centre for Health and Social Care (ICHSC).[17] The THIN dataset contained 1787 169 records on patients registered with 288 UK practices for some period between 1st January 1995 and 31st March 2006. The mean follow-up time was 5 years. The entry date for a patient was the latest of the following dates: 35th birthday, date of registration with the practice, date on which the practice computer system was installed and beginning of the study period. The first occurrence of CVD for each patient was identified using Read Codes referring to myocardial infarction, CHD, stroke and transient ischaemic attacks. There was no external validation of CVD diagnosis (eg, from linkage data) and hence misclassification is possible. The censor outcome date was the earliest of: the date of the first occurrence of CVD; date of death; date of last upload of data from the practice, date they left the practice; and end of study period (31st March 2006).

The ICHSC identified the relevant variables for the study as required by the different models including age, sex, smoking status (current or not current smoker), systolic blood pressure (SBP), total serum cholesterol, HDL-cholesterol, body mass index, recorded family history of CHD in first degree relative under 60 years, diagnosis of left ventricular hypertrophy, area measure of deprivation (Townsend quintile), treatment with antihypertensive agents (β blockers, thiazides, ACE inhibitors or calcium channel blockers), aspirin and statins.

The THIN dataset available to us was almost identical to the dataset used to validate QRISK by the original QRISK team[15] and more recently by Collins and Altman.[16] The dataset we were given did not include an additional family history dataset which was made available to both the original and independent validations of QRISK. In the QRISK validations, with the help of the additional dataset, 3.9% of the included records had recorded family history of CHD whereas in our validation this figure decreased to 3.7% (3.4% men, 4.0% women). In contrast, in the SHHEC 27.4% men and 32.6% women from the baseline population are reported to have a family history of heart disease, as obtained by answers to a questionnaire which allowed for negative and positive answers. The THIN family history variable has a positive value if the patient has 'recorded' family history of CVD in first degree relative under the age of 60. A negative value represents uncertainty and can only be interpreted as a lack of recorded information instead of as a negative value for family history. All patients with a negative value must be treated as having unknown family history. Therefore, in comparison to the SHHEC under-recording of family history in THIN appears to be an issue.

Exclusion criteria

We set out to follow the initial validation of QRISK,[15] by excluding individuals less than 35 and greater than 74 years old on entry date; those with CVD or diabetes at entry date; those with invalid dates; those taking statins at entry date; and those who had missing Townsend scores.

We encountered some problems with the implementation of the exclusion criteria as the original QRISK validation did not always explicitly define it. For example, the QRISK validation excluded patients with 'invalid dates' but did not explicitly define those. Our implementation of the 'invalid dates' exclusion criteria removed records with negative time at risk, those associated with any dates which preceded the date of birth of the patient and those whose clinical measurement dates or antihypertensive/statin dates were more than 15 years prior to the entry date.

The discrepancies in the implementation of exclusion criteria led to some differences in the number of records included in the final analysis (see table 1). However, Collins and Altman produced a detailed comparison of our exclusion criteria[18] (obtained from a previous unpublished report) with that used in the validation of QRISK. They concluded that the small discrepancies will not have an important effect on any subsequent analysis especially given the size of the dataset.

The number of records used in our study after exclusions was 1 072 289 of which 529 506 (49.4%) were male (see table 1). These corresponded to 2 600 713 and 2 752 186 years of observation for men and women respectively. The number (and percentage) of patients both male and female available to the study at yearly intervals of follow-up is given in table 2.

The percentage of males reported by the other studies was similar: 46.5% for the Anderson Framingham study[2]; 44.4% for the Cox Framingham study[11]; 49.2% for the ASSIGN study[12]; and 49.6% on the cohort used for the derivation of QRISK.[14] The mean age for the participants was also similar with 48.5 for men and 49.1 for women in the Cox Framingham Study; and 48.9 for men, 48.8 for women in the ASSIGN study. The QRISK derivation cohort had a median age of 48 for men and 49 for women. The mean ages of the participants for the Anderson Framingham model were not specified, but should be similar to those of the Cox Framingham study.

Missing data

Where smoking was not recorded the patient was assumed to be a non-smoker. Missing clinical values for SBP, total and HDL cholesterol and body mass index were replaced by the mean for the sex and age-band (5 year bands) of those recorded in the THIN dataset. Where the ratio of total to HDL-cholesterol was included in the original data this was preserved. Where the ratio was missing, this was calculated from the original or imputed values as appropriate.

For both QRISK validations that used the THIN dataset, missing data were replaced with unpublished age-sex reference values from the QRESEARCH dataset used in the development of QRISK. The total serum cholesterol to high density lipoprotein ratios were replaced by reference values matched for age and sex, not the two individual components of this ratio.

Application of CVD risk models

The THIN dataset contains values of socioeconomic deprivation for each patient in the form of the Townsend quintile relevant to

Table 1 Clinical characteristics by gender for the THIN dataset published in the validation of Collins and Altman[16] and those of our own validation

	Validation paper		UEA	
	Males N = 529813	Females N = 542987	Males 529506	Females 542783
Median age (IQR)	48 (40 to 57)	49 (41 to 59)	48 (40 to 57)	49 (41to 59)
Mean systolic blood pressure mm Hg (SD)	135.6 (19.4)	132.1 (21.0)	135.6 (19.4)	132.1 (21.0)
Mean total serum cholesterol mmol (SD)	5.7 (1.1)	5.8 (1.2)	5.7 (1.1)	5.8 (1.2)
Mean HDL mmol/l (SD)	1.3 (0.4)	1.6 (0.4)	1.3 (0.4)	1.6 (0.4)
Mean total serum cholesterol/HDL ratio (SD)	4.5 (1.3)	3.9 (1.2)	4.5 (1.3)	3.9 (1.2)
Mean body mass index kg/m^2(SD)	26.6 (4.0)	26.1 (4.9)	26.6 (4.0)	26.1 (4.9)
Current smoker (%)	141113 (26.6)	124094 (22.9)	141026 (26.6)	124047 (22.9)
Family history of coronary heart disease in first degree relative under 60 years (%)	18638 (3.5)	22922 (4.2)	17973 (3.4)	21958 (4.0)
On antihypertensive treatment at entry to the cohort (%)	35066 (6.6)	56886 (10.5)	33203 (6.3) A-	52665 (9.7) A-
(A- excluding aspirin, A+ including aspirin)			35291 (6.7) A+	55125 (10.2) A+
At the time of measure of SBP			36729 (6.9) A-	57158 (10.5) A-
On ACE inhibitors at entry to the cohort	11718 (2.2)	12901 (2.4)	11628 (2.2)	12808 (2.4)
At the time of measure of SBP			13216 (2.5)	14420 (2.7)
On beta blockers at entry to the cohort	16700 (3.2)	27554 (5.1)	16549 (3.1)	27283 (5.0)
At the time of measure of SBP			17977 (3.4)	29311 (5.4)
On calcium channel blockers at entry to the cohort	9847 (1.9)	11147 (2.1)	9761 (1.8)	11035 (2.0)
At the time of measure of SBP			10684 (2.0)	12027 (2.2)
On thiazides at entry to the cohort	10630 (2.0)	23391 (4.3)	10532 (2.0)	23172 (4.3)
At the time of measure of SBP			11361 (2.2)	24730 (4.6)
On aspirin at entry to the cohort			4763 (0.9)	5586 (1.0)
Deprivation index (% in Townsend score fifths)				
1 (most affluent)	145577 (27.5)	151352 (27.9)	145439 (27.5)	151276 (27.9)
2	119132 (22.5)	125689 (23.1)	119023 (22.5)	125617 (23.1)
3	108212 (20.4)	112150 (20.7)	108140 (20.4)	112098 (20.7)
4	89290 (16.9)	90521 (16.7)	89284 (16.9)	90509 (16.7)
5 (most deprived)	67602 (12.8)	63275 (11.7)	67620 (12.8)	63283 (11.7)
No. of incident 10 year cardiovascular disease events (%)	25963 (4.9)	18027 (3.3)	26202 (4.9)	18173 (3.3)
10 year risk of cardiovascular disease events (95% CI)	9.9 (9.7 to 10.0)	6.6 (6.4 to 6.7)	9.9 (9.7 to 10.0)	6.6 (6.4 to 6.7)
Total person years of observation (years)	2603294	2753924	2600713	2752186

The figures in brackets are IQR for age and SD or percentage (%) of records for other variables as stated. The Kaplan–Meier 10-year risk estimate shows 95% CIs in brackets.

their postcode. This was converted to an equivalent Scottish Index of Multiple Deprivation[19] (needed for ASSIGN) using the values shown in table 3. ASSIGN requires an estimate of cigarettes/day for smokers, and these were imputed using mean values for each age group and sex obtained from the 2003 Scottish Health Survey.[19]

For the Cox Framingham CVD equation,[11] we considered the clinical value for SBP as treated (associated with a higher coefficient in the model) if the date of measurement of SBP was within a period of treatment with β blockers, thiazides, ACE inhibitors or calcium channel blockers. This method results in more patients, both male and female, having treated SBP than if antihypertensive treatment at entry date is considered (see table 1).

Table 2 Number of patients available, male and female, by length of follow-up in years

Years of follow-up	Females number (%)	Males number (%)
[0,1)	49266 (9.1)	53258 (10.1)
[1,2)	56535 (10.4)	57807 (10.9)
[2,3)	40538 (7.5)	41773 (7.9)
[3,4)	44475 (8.2)	45306 (8.6)
[4,5)	81459 (15.0)	78831 (14.9)
[5,6)	64257 (11.8)	62272 (11.8)
[6,7)	69111 (12.7)	65658 (12.4)
[7,8)	36103 (6.7)	33906 (6.4)
[8,9)	31813 (5.9)	29599 (5.6)
[9,10)	49825 (9.2)	44036 (8.3)
10+	19401 (3.6)	17060 (3.2)

The final dataset was input into Stata 8.1 and individual 10-year risk scores were calculated using our implementation of four models: Framingham CHD, Framingham Stroke (S), ASSIGN, and the new Cox Framingham CVD model. The risk scores predicted by the Framingham CHD and stroke models were summed (CHD+S) to give an overall Anderson Framingham probability of either disease, as recommended by the NICE guidelines.[4] The Cox Framingham score was obtained by transforming the general cardiovascular risk score produced by the equation into individual cardiovascular components for coronary heart disease and stroke by using the calibration factor for each[11] and adding the components together.

Evaluation of models

The Kaplan–Meier (K-M) product-limit estimator was used to estimate the observed risk.

For the purpose of model comparison, we looked at calibration and discrimination. We have endeavoured to produce analysis that would be easily comparable to the analysis presented in the validation of Collins and Altman[16] as a direct comparison by producing QRISK scores was not possible. Therefore, we first present calibration analysis separately for men and women by plotting mean predicted incidence by each

Table 3 Conversion of Townsend quintile deprivation value in THIN dataset to equivalent Scottish Index of Multiple Deprivation

THIN Townsend deprivation quintile	Equivalent Scottish index of multiple deprivation
1 (least deprived)	4.1
2	10.6
3	17.3
4	27.6
5 (most deprived)	60.8

model using tenths of predicted risk against observed incidence given by the K-M estimate. We also present similar mean predicted to observed (K-M) graphs using 5 year age bands. Additionally we present the ratio of predicted to observed risk for each sex and overall, where a value of 1 is indicative of good performance.

The Brier score, a measure of the accuracy calculated as the average squared deviation between predicted and observed risk, is also included. For the Brier score, a lower value represents higher accuracy.

Discrimination is the ability of the score to differentiate between people who will have an event from those who will not, over a defined period of time. We obtained summary measures of discrimination by calculating the Area Under the curve (AUROC) for each model. We also calculate the D[20] and R^2 [21] statistics, which are measures of discrimination and explained variation respectively and are specific to censored survival data. For the D measure, higher values are indicative of greater discrimination and an increase of 0.1 over other models is said to be a good marker of improved prognostic separation.

To obtain threshold measures of discrimination, patients were classified as being at high risk of CVD or not according to the current clinical threshold (ie, those with a predicted risk ≥20% over a 10-year period). CVD events occuring during the time at risk were used as the outcome to measure against.

Next, we compared the predictions of each pair of models. First, we categorised each score based on the 20% 10-year CVD risk threshold, that is, as two categories of high risk or lower risk. For the comparison, records were divided into four groups for each pair of models: those where both models agreed on a high risk prediction; those where both models agreed on a lower risk prediction and the further two groups where one model predicted high and the other one lower. For each group, the percentage of records with a CVD event and the K-M estimated incidence was calculated.

Finally, we calculated the net reclassification index (NRI)[22] as a measure of change in the risk categories assigned by the scores. The NRI was calculated as the difference in proportions moving up and down among CVD event versus non-CVD event patients. A value of say 5% would indicate that 5% more patients with a CVD event appropriately move up a category of risk than down compared with non-CVD patients.

RESULTS

The comparison of clinical characteristics for the records included in our study and in that of Collins and Altman[16] shown in table 1 indicates that the two datasets are almost identical in terms of clinical characteristics. Our inclusion criteria were more stringent, resulting in fewer records. The only notable differences are on the percentage of records with recorded family history of CHD as already discussed.

The overall K-M 10-year survival limits for men and women in the THIN dataset were 0.901 (95% CI 0.899 to 0.903) and 0.934 (CI 0.933 to 0.936) corresponding to 10-year estimated risks of 9.90% and 6.60% respectively. The crude incidence rate was 10.07 and 6.60 per 1000 person - years respectively for men and women. These are comparable to those in the database used to develop QRISK.[14]

Calibration and discrimination results

Figure 1 shows the mean predicted versus observed risk for each equation by tenths of predicted risk. For women, the plots show good calibration of the scores with some overprediction for the higher risk patients. Anderson Framingham provides the closest calibration. ASSIGN overpredicts more than both Framingham

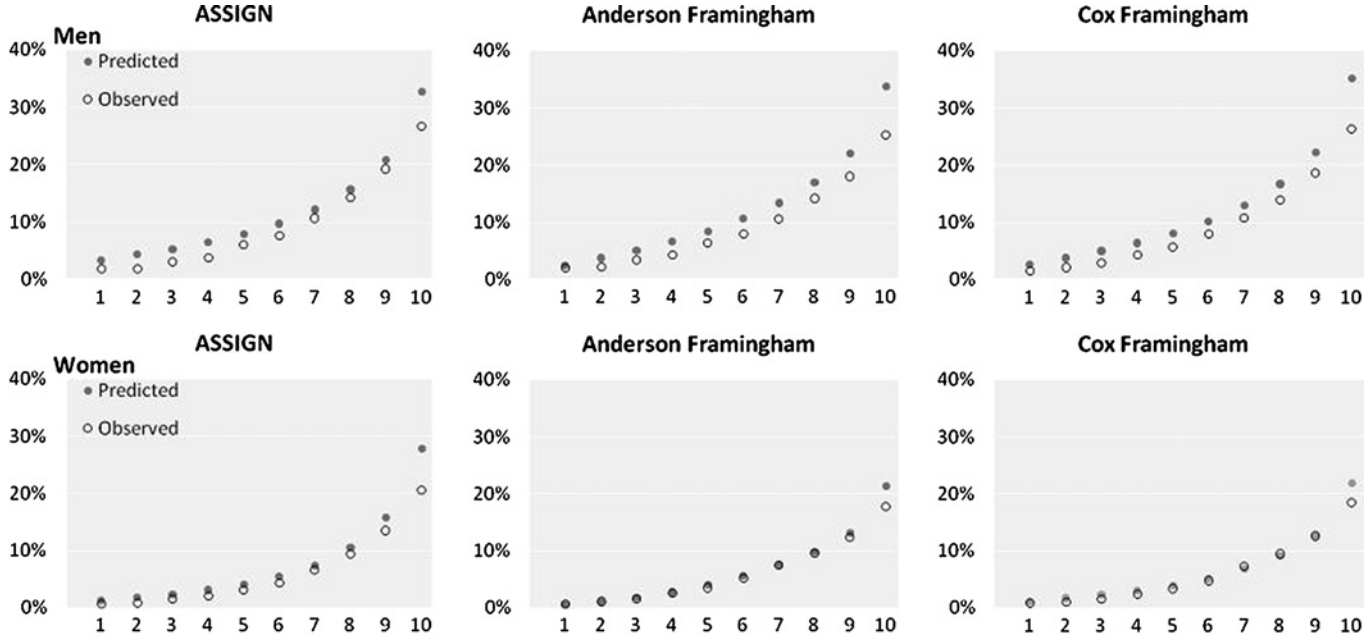

Figure 1 Predicted versus observed 10 year risk of cardiovascular disease for ASSIGN and for the Framingham equations by tenth of risk.

Table 4 Discrimination and calibration statistics for predicted 10 year risk of cardiovascular disease by ASSIGN and Framingham risk equations

	UEA results					
	ASSIGN		Anderson Framingham		Cox Framingham	
Statistic	Men	Women	Men	Women	Men	Women
AUROC	0.756	0.792	0.740	0.765	0.752	0.771
D statistic	1.35 (1.33 to 1.37)	1.58 (1.56 to 1.60)	1.26 (1.24 to 1.28)	1.39 (1.36 to 1.41)	1.32 (1.30 to 1.34)	1.41 (1.39 to 1.44)
R^2 statistic (95% CI)	30.47 (29.82 to 31.16)	37.39 (36.70 to 37.97)	27.57 (27.07 to 28.07)	31.51 (30.97 to 32.21)	29.52 (29.00 to 30.21)	32.37 (31.64 to 33.02)
Brier score	0.0517	0.0351	0.0536	0.0335	0.0535	0.0334
Predicted/observed	1.20	1.20	1.25	1.02	1.25	1.04
Predicted/observed (overall)	1.20		1.16		1.17	
	Validation paper					
	QRISK*		Anderson Framingham*		Cox Framingham*	
Statistic	Men	Women	Men	Women	Men	Women
AUROC	0.762	0.789	0.737	0.761	0.752	0.770
D statistic	1.39 (1.38 to 1.41)	1.56 (1.53 to 1.58)	1.26 (1.24 to 1.28)	1.38 (1.35 to 1.40)	1.33 (1.31 to 1.34)	1.41 (1.39 to 1.44)
R^2 statistic (95% CI)	31.71 (31.09 to 32.31)	36.64 (35.94 to 37.34)	27.31 (26.69 to 27.93)	31.18 (30.45 to 31.91)	29.52 (28.91 to 30.14)	32.32 (31.59 to 33.04)
Brier score	0.0470	0.0321	0.0545	0.0334	0.0530	0.0330
Predicted/observed	0.87	0.90	1.32	1.10	1.25	1.04
Predicted/observed (overall)	0.88		1.23		1.18	

The bottom half of the table incorporates the results of QRISK and the Framingham equations published by Collins and Altman[16] to aid comparison.
*Note that the datasets from which the results are obtained are very similar but not identical and the implementation of the Framingham risk equations may also show some differences.

equations, particularly for the higher risk women. For men, all models overpredict across all tenths of risk but this is more marked for the higher risk patients. ASSIGN provides the closest calibration but the differences are small.

Similarly, table 4 presents the ratio of predicted to observed risks for the three models for each tenth of risk. Overall, Anderson Framingham overpredicts by 16%, Cox Framingham overpredicts by 17% and ASSIGN overpredicts by 20%. For women, Anderson Framingham overpredicts by 2%, Cox Framingham overpredicts by 4%, and ASSIGN shows higher overprediction at 20%. For men, ASSIGN shows the best ratio with 20% overprediction, followed closely by both Framingham equations with 25% overprediction.

Figure 2 shows the agreement between observed risk and mean predicted risk by 5 year bands for both sexes. The diagonal indicates a perfect fit. For men, the three equations overestimate risk, with ASSIGN providing good fit for patients less than 65 years of age. For women, the Framingham equations overpredict until age 60 but then underpredict for the higher age groups. ASSIGN overpredicts across all age ranges.

We found very good agreement between our predicted to observed graphs (figure 1) for the Cox Framingham equations and those produced by Collins and Altman.[16] Some discrepancies were found for the Anderson Framingham equations and between our Framingham results (see figure 2) and those for similar figures produced by Collins and Altman.[16] In our graph, Cox Framingham overpredicts more than Anderson Framingham for the older age groups which may correspond to more patients with treated SBP (median age 59/60 years for men/women respectively).

Our summary measures of discrimination are presented in table 4. This table also includes a reproduction of the results of Collins and Altman[16] to aid comparison. In terms of our own results, from the AUROC analysis all models perform at similar levels, but ASSIGN appears to be slightly better for both men (0.756) and women (0.792). The ASSIGN Brier score is lower (more accurate) for men (0.0517) but not for women (0.0351). The *D* statistic shows ASSIGN to be the best model for both men (1.35) and women (1.58) with increases in value greater than 0.1, which are said to indicate improved diagnostic

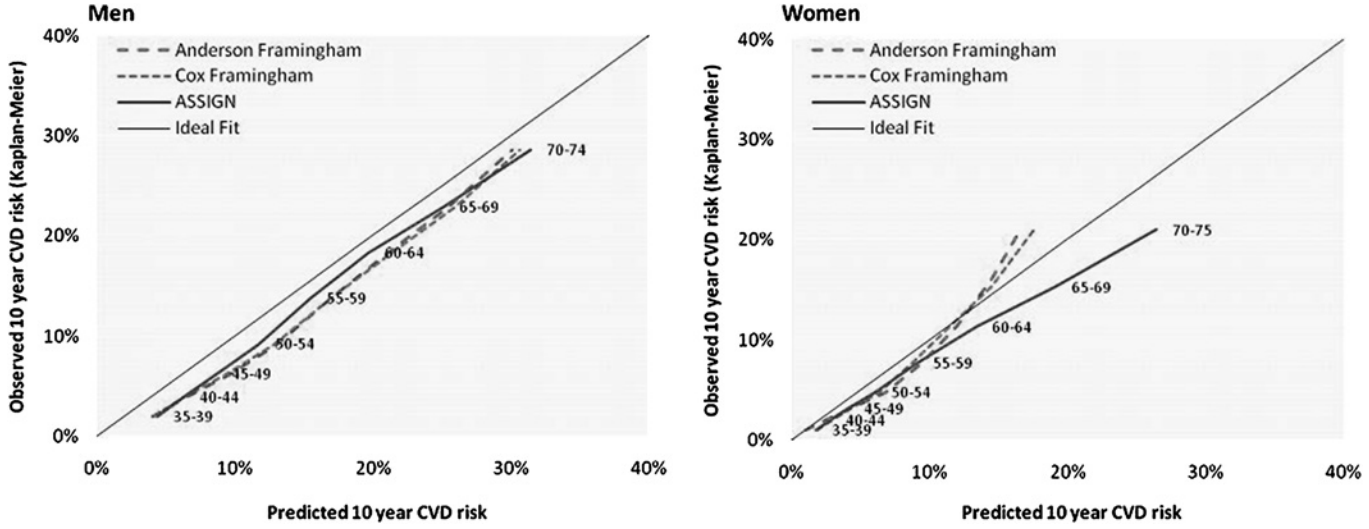

Figure 2 Predicted versus observed 10 year risk of cardiovascular disease for ASSIGN Framingham risk equations in 5 year age bands.

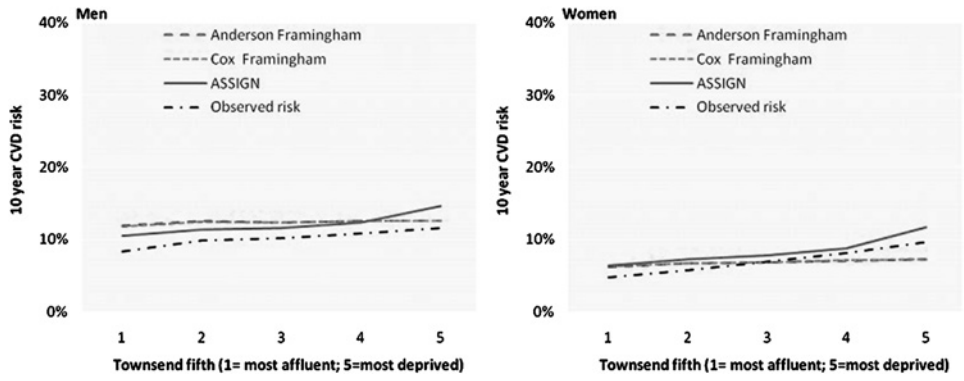

Figure 3 Predicted versus observed (Kaplan-Meier) 10 year risk of cardiovascular disease for ASSIGN and Framingham risk equations by Townsend fifth.

separation. The percentage of explained variation according to the R^2 statistic is again greater for ASSIGN (30.47% for men, 37.39% for women), followed by Cox Framingham (29.52% for men, 32.37% for women) and Anderson Framingham (27.57% for men, 31.51% for women). In comparison with QRISK, ASSIGN appears to have better discrimination for women and worse discrimination for men for the AUROC, D statistic and R^2 statistic. The Brier score gives QRISK the advantage for both men and women.

Impact of deprivation

The THIN database contains more records from the least deprived (lowest Townsend fifth) areas. Figure 3 shows a graph of mean predicted risk and observed (K-M) risk by Townsend Fifth for all models. The K-M incidence reaffirms that CVD incidence grows steadily with deprivation and the gradient of deprivation is more marked for women, where incidence nearly doubles from the least to most deprived area.

ASSIGN shows the largest deprivation gradient. Both Framingham models, which do not model deprivation explicitly, show that risk does not increase significantly with deprivation. For women, the Framingham equations under predict for the most deprived fifths.

Family history of CVD

Of the models tested, only ASSIGN takes account of family history of CVD with HRs of 1.32 and 1.63 for men and women respectively. Table 5 includes the percentage of patients from THIN with recorded family history of CVD for both men and women and the percentage of those that would be classified as high risk by each model. It also includes the K-M estimated incidence of each group for comparison. Estimated incidence is twice as high for women with a recorded family history of CVD and 3 times as high for men. ASSIGN scores place a higher proportion of those with recorded family history of CVD at high risk compared to those without history and appear to give the greatest differentiation in scores between those two groups. Both Framingham models place a higher proportion of men and women without recorded family history of CVD at high risk.

Table 5 Percentage of patients with cardiovascular disease risk score ≥20% over ten years by family history of CVD from Framingham and ASSIGN models. Observed incidence (Kaplan–Meier) for each group included for comparison

	% in dataset	Observed risk	Anderson Framingham	Cox Framingham	ASSIGN
Men					
Recorded family history of CVD	3.4	20.6	16.6	16.7	21.5
No recorded family history of CVD	96.6	6.3	18.4	18.4	16.0
Total men	100.0	9.8	18.3	18.3	16.2
Women					
Recorded family history of CVD	4.1	12.0	4.3	4.5	15.4
No recorded family history of CVD	95.9	6.3	4.5	4.9	9.3
Total women	100.0	6.6	4.5	4.9	9.5

Divergence of models

For men all models agreed with around 93% to 96% of the predictions (table 6). Both Framingham equations agreed the most with their risk predictions (79.9% agreement with lower and 16.5% with high). When models agreed with either a high or lower prediction, the K-M incidence showed that the prediction was in line with estimated incidence.

For the smaller proportion of records in which models disagreed, some models showed advantage over others according to the K-M incidence. Generally, Anderson Framingham made worse predictions than ASSIGN and Cox Framingham. ASSIGN was more in line with incidence than Cox Framingham when they disagreed but neither model may have made the correct high risk predictions given the K-M incidence.

For women, again models agreed with between 93% and 97% of predictions (table 7). The Framingham equations agreed the most with their lower risk predictions (93.3%). For their high risk predictions, ASSIGN and Cox Framingham agreed the most (4.1%). When the models agreed, the prediction was correct according to the K-M estimated incidence. When the models disagreed, Anderson Framingham appeared to be marginally more correct than Cox Framingham but both Framingham equations were less correct than ASSIGN.

The NRI of ASSIGN with respect to Anderson Framingham is 4% for men and 16% for women respectively. The NRI of ASSIGN with respect to Cox Framingham is 0% for men and 12% for women respectively. The NRI of Cox Framingham with respect to Anderson Framingham is 4% for both men and women.

DISCUSSION

The initial assessment of the THIN dataset highlighted concerns over missing values, timeliness of clinical values, time at risk and quality of recorded end points. Despite the problems highlighted, some of the characteristics of the data are reassuring. We assessed our clinical values against those on the Health survey for England[23] and found them to be comparable. Additionally, incidence of CVD recorded in THIN appears to be in line with that of the QRESEARCH dataset used for the development of QRISK and which was validated by linkage to the Office for

Table 6 Model agreement/disagreement for men

Men		Anderson Fram.(2) %	Anderson Fram.(2) % events	Anderson Fram.(2) K-M	ASSIGN(2) %	ASSIGN(2) % events	ASSIGN(2) K-M
Cox Fram. (1)	Agree low	79.9	3.0	6.3	80.1	3.1	6.4
	Agree high	16.5	12.4	22.5	14.6	13.7	24.7
	(1) Low/(2) High	1.8	11.2	20.1	1.6	11.0	18.5
	(1) High/(2) Low	1.8	16.0	27.7	3.8	9.1	15.7
Anderson Fram. (1)	Agree low				80.1	3.1	6.4
	Agree high				14.6	13.1	23.7
	(1) Low/(2) High				1.6	16.1	29.0
	(1) High/(2) Low				3.8	9.0	16.8

The tables for comparison of each pair of models show the % of records where models agree/disagree on their prediction (using a threshold of ≥20% risk over 10 year as HIGH), the crude incidence or % of CVD events and the estimated Kaplan–Meier incidence for the group.

National Statistics death certificates. Finally, this is the type of data that will be used in the systematic assessment of CVD risk proposed recently and so understanding how different risk assessment scores will behave when applied to these data is important.

The THIN dataset contained the clinical measurements, smoking and diabetes status that were recorded closest to the entry date for each variable. Examination of the dataset revealed a number of dates to which clinical measurements were attributed that were remote from the entry date. We chose an arbitrary cut-off point of 15 years to designate invalid dates and remove records. However, the relevance of clinical values that are far removed to the actual state of health of the patient at the entry date is questionable. For example, the mean time difference between entry date and measurement was 1.8 years for SBP, 3.1 years for total cholesterol, and 2.5 years for smoking status so risk assessments were not being made with the characteristics of the patient at entry date.

If data in primary care databases are used for systematic CVD risk assessment, missing data will have to be imputed, at least at present. For this, replacement values from the Health Surveys for England appeared to be a valid alternative although THIN imputation appeared to give values slightly closer to estimated incidence and was used. In the longer term, primary care databases should improve on the recording of information and reduce the amount of missing or uncertain data.

Calibration analysis showed that the Framingham models were well calibrated for women and overestimated risk for men. ASSIGN overestimated risk for men and women. The differences in calibration performance for all models were wider for the higher age groups where the incidence is higher. When looking at both men and women Anderson Framingham showed some advantage overall.

Discrimination is considered a more important component of the accuracy of a risk score (eg, Jackson[24]), as calibration can be improved for different populations. When looking at discrimination analysis against the ≥20% risk over 10 years threshold, ASSIGN had slightly better overall discrimination test results, although this depends on which test in considered.

ASSIGN showed better gradient of risk for deprivation. It is worth noting that THIN only includes quintile median values for deprivation and they had to be transformed to the equivalent Scottish Index of multiple deprivation. Also, the social gradient in risk which THIN shows appears unduly flat in comparison to that of the SHHEC, particularly for men, and has been highlighted as problematic by Tunstall-Pedoe *et al*.[25] Problems with the Townsend measure of deprivation have also been highlighted by Morris *et al*[26] who refer to it as outdated. It is possible therefore that the real gradient of deprivation in risk is even larger than that captured in THIN, and in that case ASSIGN could improve its advantage over the Framingham equations in real application.

ASSIGN showed better discrimination for both men and women with recorded family history who appear to be at a much higher risk of the disease according to the K-M incidence.

Furthermore, when looking at the agreement between the models, we found that models agreed with between 93 and 97% of the risk assessments when placing patients in a high or lower risk category and, for those cases where models do agree, the categorisation is substantiated by K-M estimates. For women, the agreement with high risk was less by different models than it was for men whereas the agreement with low risk was higher.

Table 7 Model agreement/disagreement for women

Women		Anderson Fram. (2) %	Anderson Fram. (2) % events	Anderson Fram. (2) K-M	ASSIGN(2) %	ASSIGN(2) % events	ASSIGN(2) K-M
Cox FRAM. (1)	Agree low	93.3	2.8	5.5	89.7	2.4	4.8
	Agree high	3.7	12.6	20.6	4.1	13.6	22.0
	(1) Low/(2) High	0.8	12.5	21.4	5.4	11.2	19.7
	(1) High/(2) Low	1.2	13.0	20.0	0.8	7.5	12.7
Anderson Fram. (1)	Agree low				89.7	2.4	4.8
	Agree high				3.8	13.2	21.8
	(1) Low/(2) High				5.6	11.6	19.9
	(1) High/(2) Low				0.7	9.1	15.5

The tables for comparison of each pair of models show the % of records where models agree/disagree on their prediction (using a threshold of ≥20% risk over 10 year as HIGH), the % of records in that group with a CVD event and the estimated Kaplan–Meier incidence for the group.

For those 3–7% patients where models gave a different risk category assessment, ASSIGN was closer to the K-M incidence but the differences were small.

The NRI also put ASSIGN ahead of Anderson Framingham for both men and women and of Cox Framingham for women.

In conclusion, using any of the models for initial systematic assessment of high or lower CVD risk would result in the majority of men and women to which the models apply getting very similar assessment and hence prioritisation for further investigation or treatment. In the smaller proportion of patients in which using a different model would have a different outcome, ASSIGN showed an advantage. Furthermore the application of ASSIGN would favour those in the most deprived areas and also differentiate better those with a recorded family history of CVD.

When comparing our results with those of Collins and Altman (table 4) we noticed some variations in model performance for the Framingham equations which may be due to different exclusion criteria, small changes on interpretation of the data (eg, treated SBP) and different imputation methods. For example, their ratio of predicted to observed risks put Cox Framingham above Anderson Framingham with overall ratios of 18% and 23% respectively. In our analysis Anderson Framingham shows a better ratio with 16% overprediction and Cox Framingham follows with 17% overprediction.

For some of the measures of calibration and discrimination such as AUROC, the *D* statistic and the R^2 statistic, ASSIGN performed as well as or better than QRISK. However, as the difference in performance of the models is marginal and small differences in interpretation of the clinical factors appear to have some impact in all measures of model performance, making claims about which model is best has to be done with caution. We feel that this is one of the important conclusions of this paper. All models displayed low sensitivity, particularly for women and specificity (analysis available from the authors).

The ASSIGN equation was derived from a Scottish cohort study, the Scottish Heart Health Study,[13] which recruited from 1984 to1987 at a time when the population of Scotland was experiencing higher incidence of CVD than other populations. It is therefore not surprising that it overestimates the risk when applied to a current English population which has been experiencing a decline in CVD over the last decade.[23] If calibration of the ASSIGN score to the current UK population results in improved discrimination, then ASSIGN could become the score of choice for the UK. However, as the results show, wide differences in calibration only lead to marginal differences in discrimination so recalibration may still result in marginal differences in discrimination performance between the competing models. All of the scores tested are similar (eg, 3 of them are based on the Cox proportional hazards model) and it may be that large improvements in discrimination may require a different type of model. For example, Neural Networks have been applied to the prediction of CVD with some success[27] and may be an alternative.

A new QRISK2 equation[28] has been externally validated on the THIN dataset.[29] The validation dataset was larger than the one used here, comprising patients registered from 1 January 1993 to 20 June 2008. QRISK2 is a more complex model, incorporating self assigned ethnicity and variables for other relevant conditions including rheumatoid arthritis, chronic renal disease and type 2 diabetes. QRISK2 was compared against the modified Anderson Framingham equation recommended by NICE as well as to the original QRISK equation. As the differences in performance between QRISK and QRISK2 were marginal some of the conclusions of this paper can be extended to QRISK2.

Acknowledgements We thank David Clucas, David Cracknell and Louise Thornton from the Information Centre for supplying the dataset and advice on its analysis. We thank Mary Thompson and Mustafa Dungarwalla for help with the ethical approval process. We are grateful for support to Neil Poulter from the NIHR Biomedical Research Centre funding scheme.

Competing interests None.

Ethics approval This study was conducted with the approval of the Ethical Approval. Ref. 08/H0305/2 Cambridgeshire 4 Research Ethics Committee.

Contributors Margaret Robins conducted data pre-processing and implementation of some of the risk models to obtain the different risk scores for performance analysis. Beatriz de la Iglesia supervised the project, contributed to implementation of the models, conducted the calibration and discrimination analysis and wrote the paper. Jane Skinner checked our implementation of the models, supplied advice and discrimination analysis on Stata, and contributed to the editing of the paper. Both Neil Poulter and John Potter provided clinical expertise and contributed to the editing of the paper.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Wood D,** Wray R, Poulter N, *et al*. Joint British Societies' Guidelines on Prevention of Cardiovascular Disease in Clinical Practice. *Heart* 2005;**91**(Suppl 5):1–52.
2. **Anderson KM,** Wilson PW, Odell PM, *et al*. An updated coronary risk profile—a statement for health professionals. *Circulation* 1991;**83**:356–62.
3. **Anderson KM,** Odell PM, Wilson PWF, *et al*. Cardiovascular-disease risk profiles. *Am Heart J* 1991;**121**:293–8.
4. **Cooper A,** Nherera L, Calvert N, *et al*. *Clinical Guidelines and Evidence Review for Lipid Modification: Cardiovascular Risk Assessment and the Primary and Secondary Prevention of Cardiovascular Disease*. London: National Collaborating Centre for Primary Care and Royal College of General Practitioners, 2008. http://www.nice.org.uk/CG067.
5. **Brindle P,** Beswick A, Fahey T, *et al*. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. *Heart* 2006;**92**:1752–9.
6. **Brindle P,** May M, Gill P, *et al*. Primary prevention of cardiovascular disease: a web-based risk score for seven British black and minority ethnic groups. *Heart* 2006;**92**:1595–602.
7. **Lloyd-Jones DM,** Nam BH, D'Agostino RB Sr, *et al*. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *JAMA* 2004;**291**:2204–11.
8. **Nilsson PM,** Nilsson JA, Berglund G. Family burden of cardiovascular mortality: risk implications for offspring in a national register linkage study based upon the Malmo Preventive Project. *J Intern Med* 2004;**255**:229–35.
9. **Mitchell R,** Fowkes G, Blane D, *et al*. High rates of ischaemic heart disease in Scotland are not explained by conventional risk factors. *J Epidemiol Community Health* 2005;**59**:565–7.
10. **Engstrom G,** Tyden P, Berglund G, *et al*. Incidence of myocardial Infarction in Women. A cohort study of risk factors and modifiers of effect. *J Epidemiol Community Health* 2000;**54**:104–7.
11. **D'Agostino RB Sr,** Vasan RS, Pencina MJ, *et al*. General cardiovascular risk profile for use in primary care. *Circulation* 2008;**117**:743–53.
12. **Woodward M,** Brindle P, Tunstall-Pedoe H. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN Score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 2007;**93**:172–6.
13. **Smith WC,** Crombie IK, Tavendale R, *et al*. The Scottish Heart Health Study: objectives and development of methods. *Health Bull* 1987;**45**:211–17.
14. **Hippisley-Cox J,** Coupland C, Vinogradova Y, *et al*. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;**335**:136.
15. **Hippisley-Cox J,** Coupland C, Vinogradova Y, *et al*. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;**94**:34–9.
16. **Collins GS,** Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009;**339**:b2584.
17. Ethical Approval. Ref. 08/H0305/2 Cambridgeshire 4 Research Ethics Committee.
18. **Collins GS,** Altman DG. *Report to the Department of Health: Independent Validation of QRISK on the THIN Database*. Oxford: University of Oxford, 2008.
19. Assign Cardiovascular Disease Risk Score, http://www.assign-score.com/.
20. **Royston P,** Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;**23**:723–48.
21. **Royston P.** Explained variation for survival models. *Stata Journal* 2006;**6**:1–14.
22. **Pencina MJ,** D'Agostino RB Sr, D'Agostino RB Jr, *et al*. Evaluating the added predictive ability of a new biomarker: from area under the ROC

curve to reclassification and beyond. *Stat Med* 2008;**27**:157—72; discussion 207—12.
23. **Department of Health. Health Survey for England.** http://www.dh.gov.uk/en/Publicationsandstatistics/PublishedSurvey/HealthSurveyForEngland/index.htm.
24. **Jackson R.** Cardiovascular risk prediction: are we there yet? *Heart* 2008;**94**:1—3.
25. **Tunstall-Pedoe H,** Woodward M, Watt G. ASSIGN, QRISK, and validation. *BMJ* 2009;**339**:b3514.
26. **Morris R,** Petersen I, Marston L, *et al*. Bespoke cohort studies needed. *BMJ* 2009;**339**:b3512.
27. **Voss R,** Cullen P, Schulte H, *et al*. Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks. *Int J Epidemilo* 2002;**31**:1253—62.
28. **Hippisley-Cox J,** Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;**336**:1475—82.
29. **Collins GS,** Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010;**340**: c2442, doi:10.1136/bmj.c2442.