QRISK Cardiovascular Disease Risk Prediction Algorithm – comparison of the revised and the original Analyses

Technical Supplement 1

01 November 2007

Authors:

Julia Hippisley-Cox
Carol Coupland
Yana Vinogradova
John Robson
Peter Brindle

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

1 Contents

1 Contents 1-2

2 Tables 2-3

3 Purpose of document 3-4

4 Background 4-4

5 Data quality 5-5

6 Missing data and multiple imputation 6-5

7 Aims 7-5

8 Prior reading 8-6

9 Methods 9-6

9.1 Summary of methods reported in original paper 9-6

9.2 Multiple imputation used in the original analysis 9-6

9.2.1 ICE procedure and options 9-7

9.2.2 Original ICE procedure 9-7

9.2.3 Cholesterol/HDL ratio 9-7

9.3 Multiple imputation in the revised analysis 9-7

9.3.1 ICE procedure and options 9-8

9.3.2 Total Cholesterol/HDL ratio 9-8

9.3.3 Revised ICE procedure 9-8

9.4 Sensitivity analysis 9-8

9.5 Pattern of missing values, predictions, calibration and discrimination statistics 9-10

10 RESULTS 10-11

10.1 Pattern of missing values 10-11

10.2 Comparison of summary statistics using recorded values and imputed data 10-14

10.3 Sensitivity analyses 10-17

10.4 Calibration and discrimination statistics 10-18

10.5 Proportion of patients at high risk by deprivation quintile 10-20

10.6 Age and sex 10-21

11 Summary 11-22

12 Acknowledgements 12-22

13 References 13-23

14 Appendix 1: Variable names on the QRISK stata file 14-25

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

15 Appendix 2.....15-26

16 Appendix 3.....16-29

2 Tables

Table 1 Summary of the ICE procedures used in the sensitivity analyses 9-8
Table 2 Comparison between patients with and without smoking status recorded in the derivation cohort 10-12
Table 3 Comparison between patients with and without body mass index recorded in the derivation cohort 10-13
Table 4 Comparison between patients with and without systolic blood pressure recorded in the derivation cohort 10-13
Table 5 Comparison between patients with and without cholesterol/HDL recorded in the derivation cohort 10-14
Table 6: Summary statistics for recorded and imputed data derived from the original and revised analysis (derivation cohort) 10-14
Table 7: Comparison of hazard ratios (95% CI) for original and revised models. 10-17
Table 8: Calibration and discrimination statistics for the original published in the BMJ[1] 10-18
Table 9: Calibration and discrimination statistics (with 95% confidence intervals) based on the QRESEARCH validation cohort as published in the Heart Paper[3]. 10-19
Table 11: Percentage of patients with a cardiovascular disease risk score ≥ 20% by fifth of Townsend score and sex in patients aged 35 to 74 years in the validation cohort for the original published in the BMJ[1] and revised analysis published in Heart[3]. 10-20
Table 12: Percentages of patients by age-band and sex with cardiovascular disease risk ≥20% in the validation cohort for the original published in the BMJ[1] and revised analysis published in Heart[3]. 10-21
Table 13 Adjusted hazard ratios and 95% confidence interval for risk prediction models using data imputed using different approaches (ie sensitivity analyses) 15-26
Table 14. Analysis 4 in which cholesterol/HDL was log transformed prior to inclusion in the ice procedure 15-28
Table 10: Discrimination statistics based on the QRESEARCH validation cohort – missing values in the validation cohort have been replaced using the revised approach to multiple imputation that was applied in the derivation cohort. 16-29

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

3 Purpose of document

This is a technical supplement which describes more details of the statistical methods used in the original and revised QRISK cardiovascular risk prediction algorithm published in the British Medical Journal[1 2].

The document has been written for academics with a particular interest in multiple imputation and multivariable modelling.

Technical questions regarding this document should be directed to Julia.hippisley-cox@nottingham.ac.uk or Carol.coupland@nottingham.ac.uk.

4 Background

In July 2007, we published a paper in the BMJ describing the derivation and validation of QRISK which is a new cardiovascular disease risk prediction algorithm[1].

This is a novel risk prediction algorithm which includes traditional risk factors included in the Framingham equation but also includes body mass index, family history of cardiovascular disease, social deprivation and the use of blood pressure treatment. The resulting algorithm performed well compared with Framingham in terms of discrimination and calibration and resulted in a significant reclassification of patients from high risk to low risk and vice versa.

Whilst QRISK has generally been well received, the publication of the paper sparked an important debate because of the apparent lack of a relationship between cholesterol and risk of CVD on the initial model. This debate and further discussion with colleagues (Professors Patrick Royston and Richard Peto) prompted us to undertake and publish a revised analysis on the BMJ website[2]. The revised analysis incorporated changes to the base population as patients currently prescribed statins at baseline (1% of the total population) were removed from the analysis. It also includes a change and a correction to the implementation of the multiple imputation to take account of missing data.

The revised QRISK algorithm, was then used in a second validation study designed to test the performance of QRISK in practices contributing to the THIN dataset. Practices contributing to the THIN database use a different clinical computer system from practices which contribute to the QRESEARCH database. This paper has now been published in Heart[3] and it includes a comparison of the model performance statistics from both the original QRESEARCH cohort and the THIN cohort.

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

5 Data quality

The QRESEARCH database was used for this analysis. The database has been used for a wide range of studies of cardiovascular disease[4-10] and cardiovascular risk factors[11-18], as well as the pharmacoepidemiology of anti-inflammatory drugs[19,25,26] and statins[19 20,27]. This database has been validated by comparing birth rates, death rates, consultation rates, prevalence and mortality rates with other data sources including the General Household Survey and the General Practice Research Database (GPRD)[21]. There is a good correspondence for all of these measures although in some instances QRESEARCH prevalence figures[22] of chronic diseases such as diabetes, hypertension, stroke are marginally higher than less recent data. The age-sex structure of the QRESEARCH population is similar to that reported in the UK 2001 census[21]. We have also compared practices taking part in regional research networks on these and other measures and found a good correspondence[23]. Detailed analyses have shown good levels of completeness and consistency[24].

6 Missing data and multiple imputation

Routinely collected electronic data has many advantages for research in terms of size, representativeness and generalisability. Missing data is an unavoidable problem in all clinical research especially when the research is based on electronic data collected as part of routine clinical care within general practice.

Multiple imputation is a relatively new statistical technique designed to reduce the biases which can occur in 'complete case' analysis along with a substantial loss of power and precision[22-25]. Multiple imputation allows patients with incomplete data to still be included in analyses and makes full use of all the available data, increasing power and precision[26]. The imputation technique involves creating multiple copies of the data and replaces missing values with imputed values based on suitable random sample from their predicted distribution.

7 Aims

This report has two key aims:

- To describe in detail the ICE models (Imputation by Chained Equations) used in the original and revised QRISK models and the associated sensitivity analysis so that the reader can discern the effect of different ICE models on the analysis.
- To report on the original and revised models in terms of the predictions, calibration and discrimination statistics using the original QRESEARCH validation dataset (note – these data are already in the public domain[3]).

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

8 Prior reading

This report assumes the reader is familiar with the original published paper[1], the authors' reply[2] and the second validation study[3] all of which are in the public domain.

9 Methods

9.1 Summary of methods reported in original paper

The main description of the derivation and validation of QRISK has been described in detail elsewhere[1] and is very briefly summarized here. The overall aim was to derive a new cardiovascular disease risk score (QRISK) for the United Kingdom and validate its performance against the established Framingham cardiovascular disease algorithm and a newly developed Scottish score (ASSIGN). The study was a prospective open cohort study using routinely collected data from general practice. The setting was UK practices contributing to the QRESEARCH database. The derivation cohort consisted of 1.28 million patients, aged 35-74 years registered at 318 practices between 01 Jan 1995 and 01 April 2007 free of diabetes and existing cardiovascular disease. The validation cohort consisted of 0.61 million patients from 160 different practices (a one third random sample of practices). Our endpoint was first recorded diagnosis of cardiovascular disease i.e. incident diagnosis of cardiovascular disease between 01 Jan 1995 and 01 April 2007. Cardiovascular disease includes myocardial infarction, coronary heart disease, stroke and transient ischaemic attacks. The risk factors we examined were: age, sex, smoking status, systolic blood pressure, ratio of total serum cholesterol/HDL cholesterol, body mass index, family history of coronary heart disease in first degree relative under 60 years, area measure of deprivation, existing treatment with antihypertensive agents.

9.2 Multiple imputation used in the original analysis

Initially our models were fitted using patients without any missing data (complete case analysis). However, since patients with complete data have a different health status and risk of cardiovascular disease compared with those with incomplete data, we fitted our principal models on the basis of multiple imputed datasets using Rubin's rules to combine effect estimates and estimate standard errors[27]. We used the ICE command in Stata to perform the multiple imputation[28 29.]

The use of multiple imputation makes the assumption that the missing data are "Missing At Random". "Missing at Random" generally refers to the situation where any systematic differences between the missing values and the observed values can be explained by differences in the observed data. "Missing Completely At Random" refers to the situation where there are no systematic differences between the missing values and the observed values.

In the original analysis, we used multiple imputation in the model derivation dataset to replace missing values for systolic blood pressure, body mass index, smoking, total cholesterol and

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

HDL cholesterol as all of these variables were under consideration for use in the risk prediction model.

We used the default of linear regression for imputing continuous variables within the ICE procedure and of logistic regression for the binary variable (smoking status). The default method of imputing missing values was used which samples from the posterior predictive distribution of each variable requiring imputation.

9.2.1 ICE procedure and options

We chose the following options for the ICE procedure:

- Five imputations. The choice of the number of imputations is a balance between the number needed to get robust results and practical considerations of memory size and processing power taking account of the huge size of the dataset and the limitations of the memory allocation possible in the STATA\SE.
- Set seed to 7. This was done to enable the replication of results. Seven is an arbitrary choice. The default is zero which means no seed is set and if the command was executed more than once, slightly different results would be obtained each time because of how the routine has been coded in the software.
- "Passive" option for the interaction term between systolic blood pressure and antihypertensive treatment. The "passive option" allows the use of "passive" imputation of variables that depend on other variables, some of which are imputed for example interaction terms.

9.2.2 Original ICE procedure

```
ice age hdl chol sbp bmi FH LVH smok town asp stat bptr sbpt logt
using data\CVD_imp.dta, m(5) seed(7) passive(sbpt:sbp*bptr)
```

A list of the variable names included on the analysis file can be found in Appendix 1.

9.2.3 Cholesterol/HDL ratio

- We imputed total serum cholesterol and HDL separately in the original ICE model
- We then calculated the ratio variable by dividing total serum cholesterol by HDL.

9.3 Multiple imputation in the revised analysis

Following publication of the results of the original model in the BMJ, in response to rapid responses and direct communication with other researchers which mainly focussed on the unexpected lack of association with the total cholesterol/HDL ratio and the inclusion of patients taking statins we carried out further analyses including a revised imputation of missing data. These revised imputed values were then used along with the recorded values to rerun the risk prediction model and validate the model in the validation cohort. In the revised

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

analysis we also excluded patients who were taking statins at baseline (1.1% of the cohort) as these patients are likely to have been identified as being at high risk.

9.3.1 ICE procedure and options

We again used the ICE command for the imputations with the same options as in the original analysis (i.e. 5 imputations, seed of 7 and passive option for interaction term), but we now also included the censoring indicator in the imputation model. The censoring indicator flags cardiovascular events as 1 and censoring as 0 and is denoted in Stata as _d. The _d variable had inadvertently been left out previously from the ICE procedure although the log (survival time) term had been included. This was a programming error.

As recommended by Van Burren[30], we also included additional variables which we considered might also increase the plausibility of the missing at random assumption, namely number of prescriptions for aspirin, statins and antihypertensives in follow-up period, diagnosis of hypertension and diagnosis of diabetes in follow-up.

9.3.2 Total Cholesterol/HDL ratio

As in the original procedure, we imputed total serum cholesterol and HDL separately in the revised ICE model. However, we then calculated the ratio of cholesterol/HDL based on the imputed cholesterol and HDL values. We used the patient's original ratio variable where recorded and the calculated ratio derived from the imputed values where it was missing. We then constrained the range of values for the ratio term to lie between 2 and 10 (biologically plausible values).

9.3.3 Revised ICE procedure

Addition of outcome variable (_d) and additional variables

```
ice age hdl chol sbp bmi FH LVH smok town asp stat bptr sbpt
count_aspirin flag_hypertension count_statin  count_bptreat
flag_diabetes logt _d  using data\CVD_imp_new2.dta, m(5) seed(7)
passive(sbpt:sbp*bptr)
```

9.4 Sensitivity analysis

We undertook a sensitivity analysis to help us understand the different effects of removing statin users at baseline, changing the method used to derive the cholesterol ratio term, adding the outcome variable to the multiple imputation procedure and adding two new diagnoses and number of prescriptions for statins, aspirin and antihypertensive drugs to the multiple imputation. We compared the estimates obtained in the risk prediction models using imputed data derived using a number of different specifications.

Table 1 **Summary of the ICE procedures used in the sensitivity analyses**

Sensitivity analysis	Description of sensitivity analysis

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

Original analysis in the BMJ[1]	original analysis in the BMJ paper[1] `ice age hdl chol sbp bmi FH LVH smok town asp stat bptr sbpt logt using data\CVD_imp.dta, m(5) seed(7) passive(sbpt:sbp*bptr)`
Analysis 1	Same as original analysis but with the revised method for calculating the cholesterol/HDL ratio term and also excluding patients prescribed statins at baseline: `ice age hdl chol sbp bmi FH LVH smok town asp stat bptr sbpt logt using data\test_1.dta, m(5) seed(7) passive(sbpt:sbp*bptr)`
Analysis 2	Same as analysis 1 but with the addition of the outcome variable to the ICE procedure and also excluding patients prescribed statins at baseline: `ice age hdl chol sbp bmi FH LVH smok town asp stat bptr sbpt logt _d using data\test_2.dta, m(5) seed(7) passive(sbpt:sbp*bptr)`
Analysis 3	Same as analysis 2 but with the addition of diabetes during follow-up and the hypertension variable to the ICE procedure and also excluding patients prescribed statins at baseline: `ice age hdl chol sbp bmi FH LVH smok town asp stat bptr sbpt flag_hypertension flag_diabetes logt _d using data\test_3.dta, m(5) seed(7) passive(sbpt:sbp*bptr)`
Revised analysis published by the BMJ as an authors' rapid response[2]	Same as analysis 3 but with the addition of count of statin scripts count of aspirin scripts and count of antihypertensive scripts to the ICE procedure and also excluding patients prescribed statins at baseline: `ice age hdl chol sbp bmi FH LVH smok town asp stat bptr sbpt count_aspirin flag_hypertension count_statin count_bptreat flag_diabetes logt _d using data\CVD_imp_new2.dta, m(5) seed(7) passive(sbpt:sbp*bptr)`

Additional analysis 4	In addition, we ran an imputation model where we log transformed the ratio term before including it in the ice procedure.
	
	`lnskew0 lnrati = rati1`
	
	`ice age lnrati sbp bmi FH LVH smok town asp stat bptr sbpt count_aspirin flag_hypertension count_statin count_bptreat flag_diabetes logt _d using data\CVD_imp_new3.dta, m(5) seed(7) passive(sbpt:sbp*bptr)`

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

9.5 Pattern of missing values, predictions, calibration and discrimination statistics

We have presented the pattern of the missing values for patients in the derivation dataset, and compared characteristics of patients with and without missing values for the exposures under consideration.

We applied the estimates obtained from revised model to the validation dataset and compared the original model from the paper with the revised model in terms of the proportion of patients estimated to be at high risk by age, sex and deprivation. We did not use multiple imputation to replace missing values in the validation dataset since this would be hard to apply in practice, but instead replaced missing values of continuous variables with means by age and sex obtained from the derivation dataset, and assumed patients were non-smokers when smoking status was not recorded.

We calculated the D statistic[31] and an R squared statistic derived from the D statistic[32] which are measures of discrimination and explained variation appropriate for survival models. The D statistic has been developed as a new measure of discrimination specifically for censored survival data, higher values indicate improved discrimination, and an increase in the D statistic of at least 0.1 can indicate an important difference in prognostic separation between different risk classification schemes.

We calculated the mean predicted 10 year risk for QRISK and Framingham scores, and calculated the observed 10 year risk, using the Kaplan-Meier method. We then calculated the ratios of the mean predicted 10 year risk/observed ten year risk as a measure of calibration using the validation dataset. We also calculated the ROC statistic which is commonly cited despite its shortcomings[33].

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

10 RESULTS

10.1 Pattern of missing values

We presented more details of the pattern of missing values for each prognostic factor to the BMJ during the review process. Our analysis of the associations between missing data and the other prognostic variables show a number of associations. This suggests that the data are not missing 'Completely at Random' (MCAR). It also supports the assumption that the data may be 'Missing at Random' (MAR) meaning that whether or not a particular value is missing is associated with the observed values of other variables but not with the actual unobserved value of the variable. This gives some justification for the use of multiple imputation to impute missing values which relies on the MAR assumption. However, it should be noted that it is not possible to test for MAR within a given dataset.

In the derivation dataset, smoking status was recorded in 91% of women and 84% of men, body mass index was recorded in 79% of women and 71% of men, systolic blood pressure was recorded in 91% of women and 82% of men, total serum cholesterol was recorded in 37% of women and 36% of men and HDL was recorded in 27% of women and 25% of men. Overall 24% of women and 22% of men had complete data for all risk factors used in the Cox regression model. There is no rule of thumb as to what proportion of data can be imputed in order to give reliable estimates. In an analysis reported by Moons et al[25], where missing values were present in 50-55% of patients, the imputation of missing values still yielded less biased results compared to the commonly used complete case analysis. Moons et al raised the question as to how many missing values a predictor may have and how many subjects can be imputed before the multiple imputation method also will provide invalid results and stated that to their knowledge, there are as yet no empirical studies that show the upper limit of missing values that can validly be imputed. This has been flagged as a subject for future research[25]. However Royston used multiple imputation to obtain reasonable estimates of mandible length in fetuses in which only about one quarter had recorded values[34] and in a simulation study Donders et al reported unbiased estimates using multiple imputation in a simulated dataset where 80% of non diseased subjects were given a missing value for a test whereas the diseased subjects had no missing data[35] .

As reported in the original BMJ paper[1], and presented in Table 2 below, women with missing smoking status were less likely to be taking blood pressure treatment or to have a family history of CVD than women with smoking status recorded. Women with missing body mass index were less likely to be on blood pressure treatment, less likely to smoke and less likely to have a family history of CVD than women with body mass index recorded. Women with missing systolic blood pressure measurements were less likely to be on blood pressure treatment, less likely to smoke and less likely to have a family history and were slightly younger than women with blood pressure recorded. Women with the cholesterol/HDL ratio missing were less likely to be on blood pressure treatment or have a positive family history, were more likely to be smokers, less likely to have a family history, were younger, had a lower mean systolic blood pressure and had a slightly lower mean body mass index than

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

women with blood pressure recorded. A similar pattern was observed for men. This indicates that there are associations between missingness and some of the prognostic variables which supports our use of multiple imputation. In terms of time trends, there has been an increase in the recording of smoking status over the study period but smaller changes for the other variables. People with missing data had significantly different survival rates compared with people with recorded data as shown in the following tables. For example and as reported in the BMJ paper women with a cholesterol ratio recorded had a 10 year observed risk of a cardiovascular event of 4.0% compared with 7.9% for those with missing values and for men the values were 4.9% and 10.9% respectively. Similar patterns for 10 year observed risk were found for smoking status, systolic blood pressure and body mass index with the differences bing most marked in women. This again supports the use of multiple imputation.

Table 2 Comparison between patients with and without smoking status recorded in the derivation cohort

		Women		**Men**	
		smoking status		**smoking status**	
		missing (n=59,503)	**not missing (n=586,918)**	**missing (n=101,021)**	**not missing (n=535,732)**
antihypertensive treatment	n (%)	3,253 (5.5)	76,813 (13.1)	2,752 (2.7)	50,990 (9.5)
family history of premature CVD	n (%)	452 (0.8)	77,990 (13.3)	550 (0.5)	57,158 (10.7)
age	mean (sd)	51.7 (12.3)	50.6 (11.1)	48.4 (10.9)	49.6 (10.7)
Townsend score	mean (sd)	0.1 (3.6)	-0.5 (3.4)	0.3 (3.6)	-0.4 (3.5)
body mass index	mean (sd)	25.9 (4.9)	26 (4.8)	26.3 (4.0)	26.5 (4.0)
systolic blood pressure	mean (sd)	135.0 (22.6)	132.4 (21.4)	136.2 (20.1)	135.6 (19.5)
total cholesterol/HDL ratio	mean (sd)	3.9 (1.2)	4.0 (1.3)	4.6 (1.3)	4.6 (1.3)
observed risk at 10 years	%	13.0	6.3	9.2	9.5

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

Table 3 Comparison between patients with and without body mass index recorded in the derivation cohort

		Women		men	
		missing (n=138,803)	**not missing (n=507,618)**	**missing (n=187,208)**	**Not missing (n=449,545)**
antihypertensive treatment	n (%)	10,467 (7.5)	69,599 (13.7)	7,274 (3.9)	46,468 (10.3)
smoker	n (%)	24,014 (17.3)	125,075 (24.6)	33,819 (18.1)	145,325 (32.3)
family history of premature CVD	n (%)	4,546 (3.3)	73,896 (14.6)	4128 (2.2)	53580 (11.9)
age	mean (sd)	51.1 (11.9)	50.6 (11.0)	48.6 (10.8)	49.7 (10.7)
Townsend score	mean (sd)	0 (3.5)	-0.5 (3.4)	0.1 (3.6)	-0.4 (3.5)
systolic blood pressure	mean (sd)	134.4 (22.1)	132.2 (21.3)	136.7 (19.9)	135.4 (19.5)
total cholesterol/HDL ratio	mean (sd)	4.0 (1.2)	4.0 (1.3)	4.5 (1.3)	4.7 (1.3)
observed risk at 10 years	%	9.6	6.1	10.5	9.1

Table 4 Comparison between patients with and without systolic blood pressure recorded in the derivation cohort

		Women		men	
		missing (n=61,010)	**not missing (n=585,411)**	**missing (n=116,468)**	**not missing (n=520,285)**
antihypertensive treatment	n (%)	1,690 (2.8)	78,376 (13.4)	1,597 (1.4)	52,145 (10.0)
smoker	n (%)	6,724 (11.0)	142,365 (24.3)	14,467 (12.4)	164,677 (31.7)
family history of premature CVD	n (%)	1,066 (1.7)	77,376 (13.2)	1,625 (1.4)	56,083 (10.8)
age	mean (sd)	50 (11.8)	50.7 (11.2)	47.2 (10.4)	49.9 (10.8)
Townsend score	mean (sd)	0.2 (3.6)	-0.5 (3.4)	0.3 (3.6)	-0.4 (3.5)
body mass index	mean (sd)	25.3 (4.7)	26.0 (4.8)	25.7 (3.8)	26.5 (4.0)
total cholesterol/HDL ratio	mean (sd)	3.8 (1.2)	4.0 (1.3)	4.5 (1.3)	4.6 (1.3)
observed risk at 10 years	%	13.0	6.3	10.4	9.3

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

Table 5 Comparison between patients with and without cholesterol/HDL recorded in the derivation cohort

		Women		men	
		missing (n=481,450)	**not missing (n=164,971)**	**missing (n=484,134)**	**not missing (n=152,619)**
antihypertensive treatment	n (%)	41,148 (8.5)	38,918 (23.6)	27,331 (5.6)	26,411 (17.3)
smoker	n (%)	114,133 (23.7)	34,956 (21.2)	137,894 (28.5)	41,250 (27.0)
family history of premature CVD	n (%)	47,250 (9.8)	31,192 (18.9)	34,193 (7.1)	23,515 (15.4)
age	mean (sd)	49.6 (11.3)	53.8 (10.5)	48.6 (10.8)	52 (10.2)
Townsend score	mean (sd)	-0.4 (3.4)	-0.6 (3.3)	-0.1 (3.5)	-0.8 (3.3)
body mass index	mean (sd)	25.5 (4.6)	27.1 (5.0)	26.1 (3.9)	27.3 (4.1)
systolic blood pressure	mean (sd)	129.8 (20.5)	139.5 (22.1)	133.4 (18.7)	141 (20.5)
observed risk at 10 years	%	7.9	4.0	10.9	4.9

10.2 Comparison of summary statistics using recorded values and imputed data

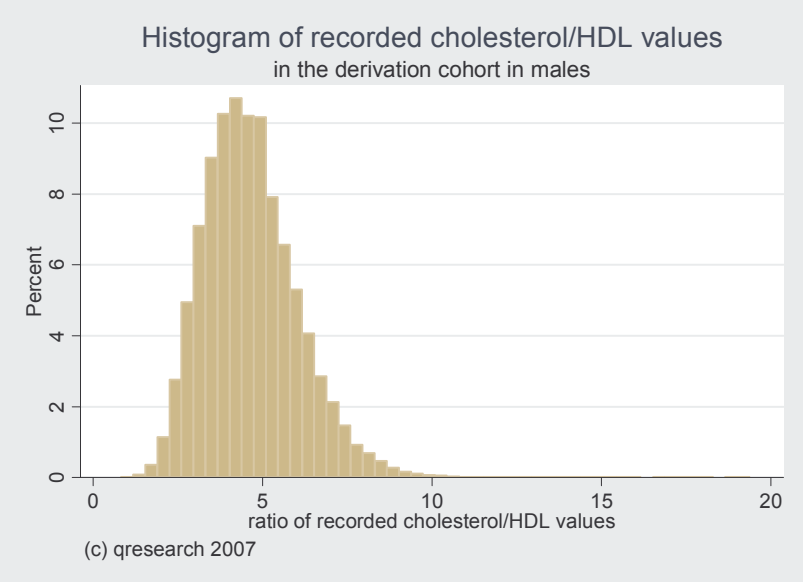
The next table shows the median (inter-quartile range) for each value based on the recorded data, the original imputation and the revised imputation. The median and inter-quartile ranges of the imputed data were very similar to those from the original data (as has been found in other studies[26] which have used multiple imputation to develop prognostic models).

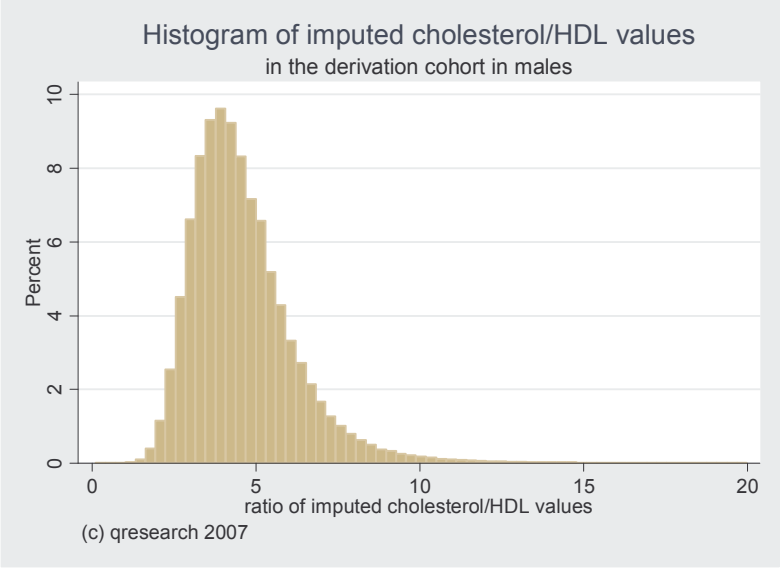
Table 6: Summary statistics for recorded and imputed data derived from the original and revised analysis (derivation cohort)

	Recorded data		**Imputed data (original method)**		**Imputed data (revised method)**	
	Median	Interquartile range	Median	Interquartile range	Median	Interquartile range
body mass index	25.6	23.0 to 28.7	25.7	23.1 to 28.8	26.0	23.7 to 28.7
systolic blood pressure	130	120 to 146	132	120 to 150	132	120 to 146
total cholesterol	5.7	5.0 to 6.5	5.7	5.0 to 6.5	5.6	4.9 to 6.4
HDL	1.4	1.1 to 1.7	1.4	1.2 to 1.7	1.3	1.1 to 1.5
cholesterol ratio	4.1	3.3 to 5.1	3.9	3.2 to 5.0	4.3	3.5 to 5.4

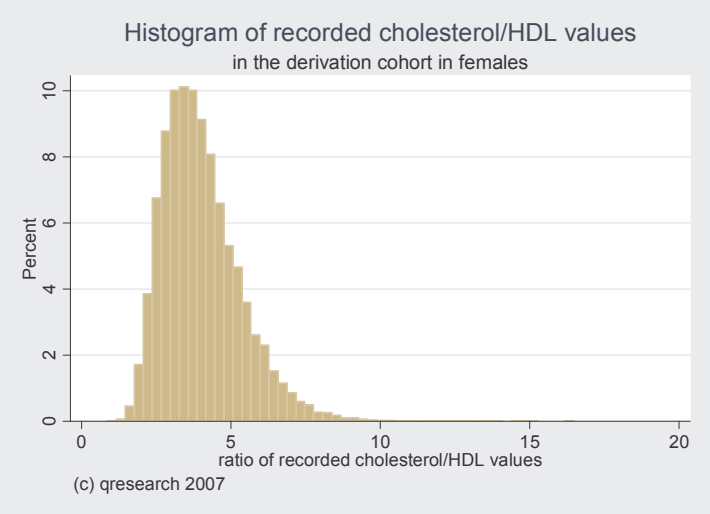
©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

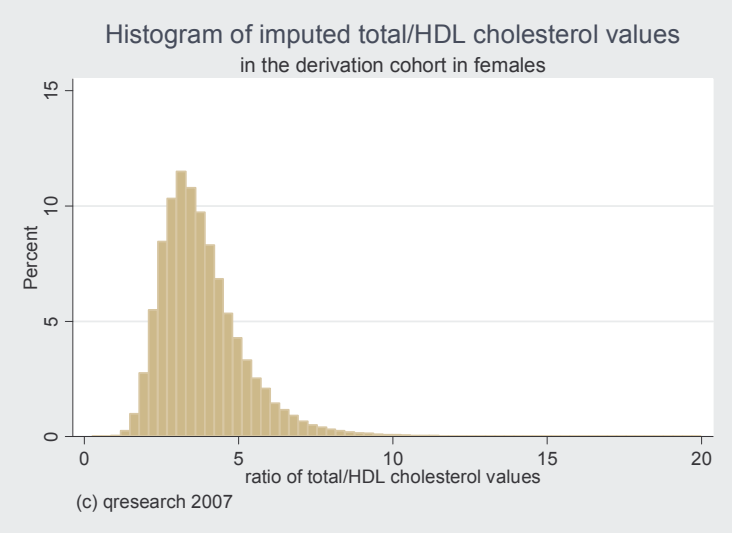
The next four figures show the histograms for the cholesterol/HDL ratio term based on the original recorded data and then from the imputed data using the revised ICE procedure (having excluded 0.07% of women and 0.1 % of males with extreme values). The distributions look similar providing reassurance that the multiple imputation generated values with similar distribution to the recorded values.





©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden





©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

10.3 Sensitivity analyses

The next table presents the adjusted hazard ratios for the original and the revised analysis. Note the units have been changed compared with the original published version for body mass index (5 unit change rather than one unit change), Townsend score (5 unit change) and systolic blood pressure (20 unit change) for ease of interpretation.

Table 7: Comparison of hazard ratios (95% CI) for original and revised models.

	Original model (published in BMJ[1])			**Revised model (BMJ authors response[2])**		
Variable	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**
Females						
Log(Age/10)	87.75	81.34	94.66	79.57	73.51	86.13
TSC/ HDL ratio (1 unit change)	1.001	0.999	1.002	1.170	1.137	1.205
Body mass index (5 unit change)	1.080	1.066	1.095	1.045	1.027	1.063
FH of premature CVD	1.229	1.187	1.273	1.209	1.166	1.253
Smoking status (current smoker)	1.530	1.487	1.574	1.531	1.482	1.583
Townsend score (5 unit change)	1.185	1.165	1.206	1.158	1.137	1.179
SBP (20 mmHg change)	1.095	1.080	1.111	1.142	1.125	1.160
BP treatment	1.734	1.674	1.796	1.709	1.646	1.774
BP treatment*SBP (20 unit change in SBP)	0.922	0.899	0.945	0.884	0.862	0.906
Males						
Log(Age/10)	50.63	47.79	53.65	56.44	52.98	60.12
TSC/ HDL ratio (1 unit change)	1.001	0.999	1.003	1.195	1.173	1.218
Body mass index (5 unit change)	1.116	1.100	1.133	1.070	1.054	1.086
FH of premature CVD	1.300	1.257	1.344	1.266	1.223	1.310
Smoking status (current smoker)	1.417	1.385	1.449	1.437	1.403	1.472
Townsend score (5 unit change)	1.090	1.074	1.107	1.088	1.072	1.105
SBP (20 mmHg change)	1.089	1.076	1.102	1.143	1.129	1.157
BP treatment	1.847	1.788	1.908	1.797	1.736	1.861
BP treatment*SBP (20 unit change in SBP)	0.874	0.852	0.896	0.841	0.818	0.863

Note: Original model included statin users at baseline but these were excluded from the revised model.

Compared with the original analysis the greatest difference in the revised analysis was for the total cholesterol/HDL ratio, which increased from 1.001 to 1.170 in women and from 1.001 to 1.195 in men, and in both cases became statistically significant. There were also slight changes in magnitude for the remaining variables.

The results from the sensitivity analyses presented in Appendix 2 show that addition of the outcome variable to the multiple imputation had the greatest effect on the estimates of hazard

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

ratios for the cholesterol ratio. This supports the work of Moons et al in 2006 showing that use of the outcome variable in the imputation of missing data is preferred[25].

The appendix also contains the results of the model based on analysis 4 (where the cholesterol ratio term was log transformed prior to including it in the ice procedure and then back transformed before inclusion in the Cox regression model). The results were very similar to the revised analysis, especially for women although there was a slight increase in the hazard ratio for the cholesterol/HDL ratio in men.

10.4 Calibration and discrimination statistics

The next table shows the calibration and discrimination statistics in the validation cohort. QRISK as published in the BMJ[1].

Table 8: Calibration and discrimination statistics for the original published in the BMJ[1]

	Original QRISK analysis (BMJ paper[1])	**Framingham analysis (BMJ paper[1])**
Females		
ROC statistic	0.788	0.774
D statistic	1.55	1.39
R squared (%)	36.4	31.7
Predicted/observed risk at 10 years	1.02	1.18
Males		
ROC statistic	0.767	0.760
D statistic	1.45	1.31
R squared (%)	33.3	29.1
Predicted/observed risk at 10 years	1.00	1.47
Predicted/observed risk at 10 years (males and females)	**1.004**	**1.35**

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

The next table shows the calibration and discrimination statistics based on the QRESEARCH validation cohort as published in the Heart paper[3]. QRISK in the revised model still performs better than Framingham for each measure. There is only a small change in these statistics between the original QRISK model and the revised model despite the change in hazard ratio for the cholesterol ratio.

Table 9: Calibration and discrimination statistics (with 95% confidence intervals) based on the QRESEARCH validation cohort as published in the Heart Paper[3].

	Revised QRISK analysis (Heart Paper[3])	**Revised Framingham analysis (Heart Paper[3])**
Females		
ROC statistic (95% CI)	0.788 (0.784 to 0.792)	0.776 (0.772 to 0.780)
D statistic (95% CI)	1.54 (1.51 to 1.56)	1.40 (1.37 to 1.42)
R squared (%) (95% CI)	36.01 (36.16 to 36.86)	31.79 (30.92 to 32.66)
Predicted/observed risk at 10 years	1.00	1.19
Males		
ROC statistic (95% CI)	0.770 (0.767 to 0.773)	0.762 (0.759 to 0.765)
D statistic (95% CI)	1.43 (1.40 to 1.45)	1.32 (1.29 to 1.34)
R squared (%) (95% CI)	32.64 (31.88 to 33.39)	29.30 (28.54 to 30.05)
Predicted/observed risk at 10 years	0.97	1.49
Predicted/observed risk at 10 years (males and females)	**0.99**	**1.36**

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

As an additional sensitivity analysis, we calculated the validation statistics having imputed the missing values in the validation dataset using the revised procedure for multiple imputation instead of replacing missing values with the age/sex reference values. We did not use multiple imputation to replace missing values in the main analysis since this would be hard to apply in practice). The results are presented and discussed in Appendix 3.

10.5 Proportion of patients at high risk by deprivation quintile

The next table shows the proportion of patients at high risk (predicted risk ≥20%) by fifth of Townsend deprivation score. Overall, a lower proportion of patients are predicted to be at high risk using the revised QRISK model compared with the original one which may in part reflect the removal of patients prescribed statins at baseline from the analysis, since this is the only difference for the Framingham analysis where a slightly lower proportion were predicted to be at high risk according to the revised analysis. The gradient between the most affluent and the most deprived fifths is very similar between the two QRISK analyses.

Table 10: Percentage of patients with a cardiovascular disease risk score ≥ 20% by fifth of Townsend score and sex in patients aged 35 to 74 years in the validation cohort for the original published in the BMJ[1] and revised analysis published in Heart[3].

Fifth of Townsend score	**Original QRISK analysis (BMJ paper)**	**Framingham analysis (BMJ paper)**	**Revised QRISK analysis (Heart Paper[3])**	**Revised Framingham analysis (Heart Paper[3])**
Females (%)				
Townsend Q1	2.97	4.56	2.75	4.51
Townsend Q2	3.84	4.81	3.55	4.74
Townsend Q3	4.88	5.29	4.64	5.20
Townsend Q4	7.47	6.31	6.83	6.18
Townsend Q5	9.93	6.31	8.67	6.21
Males (%)				
Townsend Q1	9.60	20.51	9.34	20.30
Townsend Q2	10.24	20.34	9.85	20.09
Townsend Q3	10.89	20.18	10.55	19.96
Townsend Q4	12.40	20.63	11.96	20.42
Townsend Q5	12.59	19.54	12.10	19.32
Total	**8.45**	**12.79**	**7.99**	**12.63**

Notes:

quintile 1 is the most affluent and quintile 5 the most deprived fifth
patients on statins at baseline have been removed from the revised analysis.

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

10.6 Age and sex

The next table gives a breakdown of patients at high risk by age and sex in the original and the revised analysis. In patients aged 65-74, a lower proportion of patients are at high risk using the revised QRISK model compared with the original one, however below the age of 65 there is a slight increase in the proportion of patients at high risk using the revised model compared with the original QRISK analysis.

Table 11: Percentages of patients by age-band and sex with cardiovascular disease risk ≥20% in the validation cohort for the original published in the BMJ[1] and revised analysis published in Heart[3].

Ageband	Original QRISK analysis (BMJ paper)	Framingham analysis (BMJ paper)	Revised QRISK analysis (Heart Paper[3])	Revised Framingham analysis (Heart Paper[3])
Females (%)				
35 to 44 years	0.00	0.01	0.00	0.01
45 to 54 years	0.00	0.80	0.02	0.79
55 to 64 years	2.07	7.41	2.23	7.33
65 to 74 years	34.49	24.09	31.43	24.1
Males (%)				
35 to 44 years	0.00	0.30	0.00	0.29
45 to 54 years	0.28	7.84	0.48	7.76
55 to 64 years	12.75	40.86	13.65	40.75
65 to 74 years	72.90	86.03	68.90	86.25
All patients 35-74 years	**8.45**	**12.79**	**7.99**	**12.63**

Note: patients on statins at baseline have been removed from the revised analysis.

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

11 Summary

In this report, we have presented the details of the methods used for multiple imputation to derive the original and revised QRISK cardiovascular disease risk prediction algorithms.

The sensitivity analysis presented here clearly shows the effect on hazard ratios of including the additional variables in the ICE model with the censoring indicator, in particular, having a substantial effect but only on the hazard ratio for the cholesterol ratio.

Whilst the revised model has an improved face validity with respect to the hazard ratio for cholesterol, it had minimal effect on the validation statistics and proportion of patients with a CVD risk in excess of 20%.

12 Acknowledgements

We acknowledge the contribution of Professor Patrick Royston who provided advice on application on the ICE procedure and reviewed this manuscript. All the analyses in this report have been undertaken by Julia Hippisley-Cox, Carol Coupland and Yana Vinogradova.

.

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

13 References

1. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007:bmj.39261.471806.55.
2. Hippisley-Cox J CC, Vinogradova Y, Robson J, May M, Brindle P. QRISK: Authors Response. http://www.bmj.com/cgi/eletters/335/7611/136: British Medical Journal, 2007.
3. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. The performance of the QRISK cardiovascular risk prediction algorithm in an external UK sample of patients from general practice: a validation study. *Heart* 2007:hrt.2007.134890.
4. Hippisley-Cox J, Vinogradova Y, Coupland C, Heaps M. Quality and Outcomes Time Series Analysis in QRESEARCH 2001 to 2006. *QRESEARCH research highlights*. Leeds: The Information Centre, 2007:39.
5. Hippisley-Cox J, Coupland C, Parker C, Vinogradova Y. Inequalities in the identification and treatment of hyperlipidaemia in coronary heart disease in patients with serious mental health problems: cross-sectional study. *Heart* 2007 (in press).
6. Hippisley-Cox J, Coupland C. Effect of combinations of drugs on all cause mortality in patients with ischaemic heart disease: nested case control analysis. *British Medical Journal* 2005;330:1059-1063.
7. Hippisley-Cox J, Coupland C. The effect of statins in the mortality of patients with ischaemic heart disease: population based cohort study with nested case control analysis. *Heart* 2005.
8. Hippisley-Cox J, Fenty J, Langford G, Pringle M, Coupland C. Quality of care for stroke and TIA in general practice using the new GMS contract indicators. Report to the National Audit Office. London: National Audit Office, 2005.
9. Hippisley-Cox J. Trends and variations in GMS indicators for coronary heart disease. Report to the Department of Health. Nottingham: University of Nottingham, 2005.
10. Fleming D, Elliott C, Pringle M, Anderson J, Hebbrecht G, Nardi R, et al. Electronic Health Indicator Data (eHID). *Report to the European Commission*. Birmingham: Royal College of General Practitioners, 2008 (in press).
11. Hippisley-Cox J, Pringle M. Prevalence, Care and Outcomes for patients with diet controlled diabetes in general practice: cross sectional survey. *Lancet* 2004;364:423-428.
12. Hippisley-Cox J, Ryan R. Diabetes in the UK: analysis of QRESEARCH data. Report to the Department of Health. Nottingham: University of Nottingham, 2004.
13. Hippisley-Cox J, Pringle M. Trends and variations in GMS indicators for diabetes. Report to the Department of Health. Nottingham: University of Nottingham, 2004.
14. Hippisley-Cox J, Ryan R. Incidence, prevalence and mortality of diabetes in the UK 1994 to 2003. Report to the Statistics Division, Department of Health. Nottingham: University of Nottingham, 2004.
15. Hippisley-Cox J. Obesity in the United Kingdom. Report to the Department of Health. Nottingham: University of Nottingham, 2004.

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

16. Hippisley-Cox J. The population prevalence and inter-practice variation of diabetes: analysis of QRESEARCH data. Report to the Department of Health. Nottingham: University of Nottingham, 2004.
17. Wilson A, Hippisley-Cox J, C C, Coleman T, Britton J, Barratt S. Smoking cessation treatments within primary care. Prosepctive cohort study. *Tobacco control* 2005;14:242-246.
18. Holt T, Stables D, Hippisley-Cox J, O'Hanlon S, Majeed A. Identifying patients with undiagnosed diabetes: cross-sectional survey of 3.6 million patients' electronic records. *British Medical Journal* 2007 (submitted Nov).
19. Vinogradova Y, Hippisley-Cox J, Coupland C, Logan R. Risk of colorectal cancer in patients prescribed statins, nonsteroidal anti-inflammatory drugs, and cyclooxygenase-2 Inhibitors: nested case-control study. *Gastroenterology* 2007;133:393–402.
20. Do Statins Affect Risk of Pneumonia in the General Population: Nested Case Control Study. 36th NAPCRG Annual Meeting; North American Primary Care Research Group; 2007 October 20-23rd; Vancouver.
21. Hippisley-Cox J, Fenty J, Heaps M. Trends in Consultation Rates in General Practice 1995 to 2006: Analysis of the QRESEARCH database. *QRESEARCH research highlights*. Leeds: The Information Centre, 2007:29.
22. Schafer J, Graham J. Missing data: our view of the state of the art. *Psychological Methods* 2002;7:147-177.
23. Group TAM. Academic Medicine: problems and solutions. *British Medical Journal* 1989;298:573-579.
24. Steyerberg EW, van Veen M. Imputation is beneficial for handling missing data in predictive models. *Journal of Clinical Epidemiology* 2007;60(9):979.
25. Moons KGM, Donders RART, Stijnen T, Harrell FJ. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* 2006;59(10):1092.
26. Clark T, Altman D. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology* 2003;56:28-37.
27. Rubin DB. *Multiple Imputation for Non-response in Surveys.* New York: John Wiley, 1987.
28. Royston P. Multiple imputation of missing values. *Stata Journal* 2004;4(3):227-241.
29. Royston P. Multiple imputation of missing values: Update of ice. *Stata Journal* 2005;5(4):527-536.
30. Van Buuren S, Boshuizen H, Knook D. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999;18:681-694.
31. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Statistics in Medicine* 2004;23:723-748.
32. Royston P. Explained variation for survival models. *The Stata Journal* 2006;6:1-14.
33. Cook N. Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction. *Circulation* 2007;115:928-935.
34. Royston P. Multiple imputation of missing values: update. *Stata Journal* 2005;5:1-14.
35. Donders A, van der Heijden G, Stijnend T, Moons K. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 2006;59:1087-1091.

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

14 Appendix 1: Variable names on the QRISK stata file

Variable name	Description
hdl	High density lipoprotein(mmol/l)
chol	Total serum cholesterol (mmol/l)
rati	Total serum cholesterol/HDL ratio
sbp	Systolic blood pressure (mm Hg)
bmi	Body mass index (kg/m2)
FH	Family history (yes/no)
LVH	Left ventricular hypertrophy (yes/no)
smok	Smoking status (smoker/not smoker)
town	Townsend score at output area (continuous)
asp	On aspirin at baseline (yes/no)
stat	On statins at baseline(yes/no)
bptr	On antihypertensive treatment at baseline(yes/no)
sbpt	Systolic blood pressure * antihypertensive treatment at baseline(yes/no) interaction term
age	Age at entry to cohort
logt	Log of survival time
_d	censoring indicator from stset(yes/no)
flag_hypertension	Diagnosis of hypertension at any time prior to end of study period
flag_diabetes	Incident diagnosis of diabetes during study period (yes/no)
count_statin	Number of prescriptions of statins during study period (no prescriptions coded as zero)
count_aspirin	Number of prescriptions for aspiring during study period(no prescriptions coded as zero)
count_bptreat	Number of prescriptions for antihypertensive treatment during study period

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

15 Appendix 2

Table 12 Adjusted hazard ratios and 95% confidence interval for risk prediction models using data imputed using different approaches (ie sensitivity analyses)

	Original analysis (published in BMJ[1])			Analysis 1			Analysis 2			Analysis 3		
variable	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**
Females												
Log(Age/10)	87.75	81.34	94.66	83.78	77.51	90.54	77.96	72.02	84.4	79.17	73.21	85.61
TSC to HDL ratio (1 unit change)	1.001	0.999	1.002	1.032	1.020	1.045	1.157	1.141	1.173	1.153	1.142	1.164
body mass index (5 unit change)	1.080	1.066	1.095	1.075	1.061	1.090	1.045	1.029	1.061	1.049	1.034	1.065
FH of premature CVD	1.229	1.187	1.273	1.233	1.189	1.277	1.216	1.173	1.26	1.213	1.171	1.257
smoking status	1.530	1.487	1.574	1.529	1.485	1.574	1.530	1.488	1.573	1.537	1.494	1.582
Townsend score (5 unit change)	1.185	1.165	1.206	1.177	1.157	1.199	1.157	1.136	1.178	1.16	1.139	1.181
SBP (20 mmHg change)	1.095	1.08	1.111	1.122	1.106	1.139	1.141	1.124	1.159	1.142	1.125	1.159
BP treatment	1.734	1.674	1.796	1.716	1.654	1.78	1.704	1.641	1.769	1.713	1.651	1.777
BP treatment*SBP (20 unit change)	0.922	0.899	0.945	0.905	0.882	0.928	0.894	0.871	0.917	0.885	0.863	0.907
Males												
Log(Age/10)	50.63	47.79	53.65	52.42	49.4	55.63	54.98	51.68	58.5	55.51	52.26	58.96
TSC to HDL ratio (1 unit change)	1.001	0.9991	1.003	1.058	1.049	1.067	1.180	1.166	1.195	1.181	1.167	1.195
body mass index (5 unit change)	1.116	1.100	1.133	1.093	1.076	1.110	1.060	1.044	1.075	1.076	1.06	1.092
FH of premature CVD	1.300	1.257	1.344	1.294	1.251	1.339	1.278	1.236	1.323	1.268	1.226	1.313
smoking status	1.417	1.385	1.449	1.407	1.375	1.44	1.434	1.401	1.468	1.435	1.404	1.466
Townsend score (5 unit change)	1.090	1.074	1.107	1.088	1.071	1.105	1.085	1.068	1.102	1.088	1.072	1.105
SBP (20 mmHg change)	1.089	1.076	1.102	1.105	1.091	1.118	1.137	1.123	1.151	1.141	1.124	1.158
BP treatment	1.847	1.788	1.908	1.785	1.725	1.846	1.789	1.728	1.851	1.796	1.736	1.858
BP treatment*SBP (20 unit change)	0.874	0.852	0.896	0.875	0.852	0.899	0.859	0.837	0.881	0.844	0.822	0.866

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

	Revised analysis[2]			Complete case analysis		
variable	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**
Females						
Log(Age/10)	79.57	73.51	86.13	36.87	30.26	44.92
TSC to HDL ratio (1 unit change)	1.170	1.137	1.205	1.199	1.174	1.224
body mass index (5 unit change)	1.045	1.027	1.063	1.054	1.022	1.087
FH of premature CVD	1.209	1.166	1.253	1.405	1.309	1.507
smoking status	1.531	1.482	1.583	1.383	1.292	1.481
Townsend score (5 unit change)	1.158	1.137	1.179	1.169	1.120	1.221
SBP (20 mmHg change)	1.142	1.125	1.160	1.055	1.018	1.093
BP treatment	1.709	1.646	1.774	1.593	1.486	1.708
BP treatment*SBP (20 unit change)	0.884	0.862	0.906	0.943	0.891	0.998
Males						
Log(Age/10)	56.44	52.98	60.12	30.15	25.69	35.39
TSC to HDL ratio (1 unit change)	1.195	1.173	1.218	1.248	1.226	1.270
body mass index (5 unit change)	1.070	1.054	1.086	1.043	1.009	1.078
FH of premature CVD	1.266	1.223	1.310	1.468	1.374	1.569
smoking status	1.437	1.403	1.472	1.345	1.273	1.422
Townsend score (5 unit change)	1.088	1.072	1.105	1.16	1.116	1.205
SBP (20 mmHg change)	1.143	1.129	1.157	1.074	1.043	1.106
BP treatment	1.797	1.736	1.861	1.681	1.576	1.792
BP treatment*SBP (20 unit change)	0.841	0.818	0.863	0.914	0.865	0.967

Analysis 1 used a revised method for deriving the total cholesterol/HDL ratio and excluded patients prescribed statins at baseline.
Analysis 2 was the same as analysis 1 but also included the outcome variable in the multiple imputation
Analysis 3 was the same as analysis 2 but also included hypertension and diabetes in the multiple imputation
Revised analysis was the same as analysis 3 but also included counts of aspirin, statin and antihypertensive prescriptions in the multiple imputation
Complete case analysis only included patients with recorded data for all variables in model, and excluded patients on statins at baseline.

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

Table 13. Analysis 4 in which cholesterol/HDL was log transformed prior to inclusion in the ice procedure

	Analysis 4 (lnskew)		
variable	**adjusted hazard ratio**	**Lower 95% CI**	**Upper 95% CI**
Females			
Log(Age/10)	79.27	73.326	85.695
TSC to HDL ratio (1 unit change)	1.169	1.145	1.194
body mass index (5 unit change)	1.053	1.036	1.07
FH of premature CVD	1.208	1.165	1.251
smoking status	1.537	1.494	1.581
Townsend score (5 unit change)	1.16	1.139	1.181
SBP (20 mmHg change)	1.141	1.124	1.157
BP treatment	1.707	1.646	1.771
BP treatment*SBP (20 unit change)	0.886	0.864	0.908
Males			
Log(Age/10)	57.863	54.362	61.59
TSC to HDL ratio (1 unit change)	1.232	1.205	1.259
body mass index (5 unit change)	1.072	1.055	1.089
FH of premature CVD	1.259	1.217	1.303
smoking status	1.424	1.391	1.458
Townsend score (5 unit change)	1.09	1.073	1.107
SBP (20 mmHg change)	1.14	1.126	1.154
BP treatment	1.789	1.729	1.852
BP treatment*SBP (20 unit change)	0.842	0.82	0.865

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

16 Appendix 3

We undertook an additional analysis in the validation cohort using multiple imputation to replace missing values instead of using age-sex reference values. We used the latter approach as our main analysis since this is what can be done in clinical practice where the algorithm is likely to be used and where multiple imputation would not be possible.

The results are shown in the table below. The D statistic, ROC statistic and R^2 show that QRISK has better discrimination compared with Framingham. Also as expected the discrimination statistics are improved when multiple imputation is used to replace missing values in the validation data as compared to replacement with age-sex reference values. This is likely to be because more information is used about each person so imputed values are likely to be closer to the true values.

Table 14: Discrimination statistics based on the QRESEARCH validation cohort – missing values in the validation cohort have been replaced using the revised approach to multiple imputation that was applied in the derivation cohort.

	Revised QRISK analysis (with multiple imputation also used in validation cohort)	**Revised Framingham analysis (with multiple imputation also used in validation cohort)**
Females		
ROC statistic	0.7964	0.7836
D statistic (95% CI)	1.64 (1.61 to 1.68)	1.51 (1.47 to 1.54)
R squared (%) (95% CI)	39.1 (38.1 to 40.1)	35.2 (34.0 to 36.3)
Males		
ROC statistic	0.7800	0.7680
D statistic (95% CI)	1.55 (1.51 to 1.59)	1.43 (1.39 to 1.46)
R squared (%) (95% CI)	36.4 (35.4 to 37.5)	32.7 (31.6 to 33.8)

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden

©QRESEARCJH. This document is not to be reproduced in whole or in part without the prior written permission of the author. The information contained herein is confidential and to be used only for the purpose for which it was submitted. Any use or distribution for commercial purposes is expressly forbidden