

## RESEARCH

# Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study

 OPEN ACCESS

Julia Hippisley-Cox *professor of clinical epidemiology and general practice*<sup>1</sup>, Carol Coupland *associate professor and reader in medical statistics*<sup>1</sup>, Peter Brindle *research and development programme director*<sup>2</sup>

<sup>1</sup>Division of Primary Care, University Park, Nottingham NG2 7RD, UK; <sup>2</sup>Avon Primary Care Research Collaborative, Bristol Clinical Commissioning Group, Bristol BS1 3NX, UK

## Abstract

**Objective** To develop and validate a risk algorithm (QStroke) to estimate risk of stroke or transient ischaemic attack in patients without prior stroke or transient ischaemic attack at baseline; to compare (a) QStroke with CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASC scores in patients with atrial fibrillation and (b) the performance of QStroke with the Framingham stroke score in the full population free of stroke or transient ischaemic attack.

**Design** Prospective open cohort study using routinely collected data from general practice during the study period 1 January 1998 to 1 August 2012.

**Setting** 451 general practices in England and Wales contributing to the national QResearch database to develop the algorithm and 225 different QResearch practices to validate the algorithm.

**Participants** 3.5 million patients aged 25-84 years with 24.8 million person years in the derivation cohort who experienced 77 578 stroke events. For the validation cohort, we identified 1.9 million patients aged 25-84 years with 12.7 million person years who experienced 38 404 stroke events. We excluded patients with a prior diagnosis of stroke or transient ischaemic attack and those prescribed oral anticoagulants at study entry.

**Main outcome measures** Incident diagnosis of stroke or transient ischaemic attack recorded in general practice records or linked death certificates during follow-up.

**Risk factors** Self assigned ethnicity, age, sex, smoking status, systolic blood pressure, ratio of total serum cholesterol to high density lipoprotein cholesterol concentrations, body mass index, family history of coronary heart disease in first degree relative under 60 years, Townsend deprivation score, treated hypertension, type 1 diabetes, type 2 diabetes, renal disease, rheumatoid arthritis, coronary heart disease, congestive cardiac failure, valvular heart disease, and atrial fibrillation

**Results** The QStroke algorithm explained 57% of the variation in women and 55% in men without a prior stroke. The D statistic for QStroke was 2.4 in women and 2.3 in men. QStroke had improved performance on all measures of discrimination and calibration compared with the Framingham score in patients without a prior stroke. Among patients with atrial fibrillation, levels of discrimination were lower, but QStroke had some improved performance on all measures of discrimination compared with CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASC.

**Conclusion** QStroke provides a valid measure of absolute stroke risk in the general population of patients free of stroke or transient ischaemic attack as shown by its performance in a separate validation cohort. QStroke also shows some improvement on current risk scoring methods, CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASC, for the subset of patients with atrial fibrillation for whom anticoagulation may be required. Further research is needed to evaluate the cost effectiveness of using these algorithms in primary care.

## Introduction

Cardiovascular disease is the leading cause of premature death and a major cause of disability in the UK.<sup>1</sup> In 2008 the UK government announced a major new initiative to reduce vascular risk,<sup>2</sup> building on guidelines from the National Institute for Health and Clinical Excellence (NICE) for lipid modification.<sup>3</sup> Risk factors for cardiovascular disease are now well established, and validated tools such as QRISK2<sup>4-8</sup> which predict risk of cardiovascular disease are included in clinical guidelines. QRISK2 is part of the Quality and Outcomes Framework to reward UK general practices for using it. It has been incorporated into all four major UK GP clinical computer systems, which cover more than 90% of UK general practice. This integration has allowed automated cardiovascular risk

assessment within the consultation to aid clinical decision making as well as risk stratification of GP practice populations to identify those patients who need further recall, assessment, and treatments.

QRISK2, which is updated annually, currently predicts risk of cardiovascular disease, defined as either coronary heart disease or stroke/transient ischaemic attack. It includes major risk factors such as age, sex, deprivation, ethnicity, smoking, systolic blood pressure, ratio of total serum cholesterol to high density lipoprotein cholesterol concentrations, body mass index, diabetes, rheumatoid arthritis, chronic renal disease, and atrial fibrillation. Atrial fibrillation is important since it is the most common cardiac rhythm disorder and it particularly predisposes to stroke. The risk of stroke in patients with atrial fibrillation can be reduced by anticoagulation, but the evidence for the effectiveness of aspirin is less clear.<sup>9</sup> Many patients with atrial fibrillation are not currently prescribed anticoagulation even though it is encouraged in the Quality and Outcomes Framework.<sup>10</sup> This is probably because of difficulties in case identification and concerns regarding the potential adverse effects of traditional anticoagulants such as warfarin.

New oral anticoagulants (factor Xa inhibitors and direct thrombin inhibitors) have a similar efficacy at reducing stroke in people with atrial fibrillation as warfarin, and they have a wider therapeutic range without the need for repeated monitoring of international normalised ratio (INR).<sup>11</sup> The most common tool for helping clinicians decide whether to initiate anticoagulation in people with atrial fibrillation is the CHADS<sub>2</sub> score,<sup>12</sup> which is a simple counting system which does not include many established risk factors and does not give an absolute risk of stroke. More recently, additional risk factors have been included in the CHA<sub>2</sub>DS<sub>2</sub>VASc score,<sup>13</sup> which is better at identifying low risk individuals for whom the risks of anticoagulation might outweigh the benefits<sup>14</sup> but doesn't give an absolute risk of stroke either.

While increasing the number of risk factors generally improves risk stratification, it also makes a risk factor counting tool such as CHA<sub>2</sub>DS<sub>2</sub>VASc increasingly cumbersome to use in everyday clinical practice. Our aim was therefore to develop and validate a new risk prediction algorithm to predict the risk of stroke and transient ischaemic attack (QStroke) that could be automatically populated by data held in the clinical record and calculated in the same way as QRISK2, thereby providing a simpler practical alternative to existing scores. We wanted to develop an algorithm that quantifies absolute risk of stroke in a way which can be communicated to patients to aid decision making. In particular, we wished to compare its performance with existing scores in a subset of patients with atrial fibrillation for whom anticoagulation should be considered.

## Methods

### Study design and data source

We conducted a prospective cohort study of a large UK primary care population using a similar method to our original analysis for predicting cardiovascular risk (QRISK2).<sup>15</sup> Version 34 of the QResearch database was used for this study ([www.qresearch.org](http://www.qresearch.org)). This is a large validated primary care electronic database containing the health records of 13 million patients registered from 676 general practices using the Egton Medical Information System (EMIS) computer system.<sup>15</sup> Practices and patients contained on the database are nationally representative<sup>16</sup> and similar to those on other primary care databases using other clinical software systems.<sup>4</sup> We included all QResearch practices in England and Wales once they had been using their current

EMIS system for at least a year (to ensure completeness of recording of morbidity and prescribing data), randomly allocating two thirds of practices to the derivation dataset and one third to the validation dataset.

### Cohort selection

We identified an open cohort of patients aged 25–84 years at the study entry date, drawn from patients registered with eligible practices between 1 January 1998 and 1 Aug 2012. We used an open cohort design, rather than a closed design, as this allows patients to enter the population throughout the whole study period rather than require registration on 1 January 1998, thus better reflecting the realities of routine general practice. We excluded registered patients with a prior recorded diagnosis of stroke or transient ischaemic attack at baseline because of the difficulty of distinguishing a new stroke from a review of an existing stroke in GP records. We excluded patients without a Townsend deprivation score related to a valid postcode.

We also excluded patients who were taking anticoagulants (as defined by chapter 2.8.2 of the *British National Formulary*) at baseline to reflect the clinical application of the tool for assessing patients with atrial fibrillation who might be suitable for anticoagulation and to permit better comparison with existing risk scores such as CHADS<sub>2</sub><sup>12</sup> and CHA<sub>2</sub>DS<sub>2</sub>VASc. The anticoagulants included warfarin, acenocoumarol, phenindione, dabigatran, rivaroxaban, and apixaban, though not all drugs were licensed at the start of the study period.

We did not exclude patients prescribed aspirin at baseline as aspirin is generally not considered to be effective at preventing stroke in patients with atrial fibrillation.<sup>9,17</sup> We did not exclude incident users of anticoagulants during follow-up in order to ensure the baseline population was representative of patients who might subsequently be prescribed anticoagulants.

For each patient we determined an entry date to the cohort, which was the latest of the following dates: 25th birthday, date of registration with the practice plus one year, date on which the practice computer system was installed plus one year, or the beginning of the study period (1 January 1998). Patients were censored at the earliest date of a stroke or transient ischaemic attack, death, deregistration with the practice, last upload of computerised data, or the study end date (1 August 2012).

### Stroke or transient ischaemic attack disease outcomes

The primary outcome measure of interest was the first recorded diagnosis of either stroke or transient ischemic attacks, excluding haemorrhagic stroke. The Read codes used for case identification on the GP computer record were those agreed and used in the Quality and Outcomes Framework for General Practice. The ICD-10 codes used for case identification on the Office for National Statistics death certificate were cerebral infarction (I63) and stroke not specified as haemorrhage or infarction (I64).

### Risk factors for stroke

We included the variables which are already included in the current version of QRISK2 (2013) or included in the CHADS<sub>2</sub><sup>12</sup> or CHA<sub>2</sub>DS<sub>2</sub>VASc scoring systems<sup>13</sup> as shown in table 1. The following variables were examined:

- Self assigned ethnicity (white/not recorded, Indian, Pakistani, Bangladeshi, other Asian, black African, black Caribbean, Chinese, other (including mixed))
- Age at study entry (years)
- Sex (male/female)

- Smoking status (non-smoker, former smoker, light smoker (<10 cigarettes/day), moderate smoker (10-19 cigarettes/day), heavy smoker ( $\geq 20$  cigarettes/day))
- Atrial fibrillation<sup>18 19</sup> (yes/no).
- Systolic blood pressure<sup>20</sup> (continuous)
- Ratio of total serum cholesterol to high density lipoprotein (HDL) cholesterol<sup>20</sup> (continuous)
- Body mass index<sup>15</sup> (continuous)
- Family history of coronary disease in first degree relative <60 years old<sup>15</sup> (yes/no)
- Townsend deprivation score<sup>15</sup> (output area level 2001 census data evaluated as a continuous variable)
- Treated hypertension<sup>15</sup> (diagnosis of hypertension and at least one current prescription of at least one antihypertensive agent)
- Rheumatoid arthritis<sup>21</sup> (yes/no)
- Chronic renal disease<sup>22</sup> (yes/no)
- Type 1 diabetes<sup>18</sup> (yes/no)
- Type 2 diabetes<sup>20</sup> (yes/no)
- Coronary heart disease (yes/no)
- Congestive cardiac failure (yes/no)
- Valvular heart disease (yes/no).

We restricted all values of these variables to those recorded in the person's electronic healthcare record before baseline, except for ethnicity, smoking status, systolic blood pressure, total serum cholesterol:HDL cholesterol ratio, and body mass index, where we used the values recorded closest to the study entry date and recorded before the patient had the outcome or was censored. We imputed missing values where necessary as described below.

## Model derivation and development

As in previous analyses,<sup>5</sup> we used the Cox proportional hazards model in the derivation dataset to estimate the coefficients and hazard ratios associated with each potential risk factor for the first ever recorded diagnosis of stroke or transient ischaemic attack for males and females separately. We used fractional polynomials to model non-linear risk relationships with age and body mass index where appropriate.<sup>23</sup> We tested for interactions between each variable and age and included significant interactions in the final model where they improved model fit. Continuous variables were centred for analysis. Our main analyses used multiple imputation to replace missing values for systolic blood pressure, total cholesterol:HDL cholesterol ratio, smoking status, and body mass index. Our final model was fitted based on five multiply imputed datasets using Rubin's rules to combine effect estimates and standard errors to allow for the uncertainty due to imputing missing data.<sup>24</sup> We took the log of the hazard ratio for each variable from the final model and used these as weights for the new stroke risk equations. We combined these weights with the baseline survivor function evaluated at 10 years centred on the means of continuous risk factors to derive a risk equation for 10 years' follow-up.

We conducted a sensitivity analysis in which patients with atrial fibrillation who were prescribed anticoagulants during follow-up were censored on the date of first prescription of anticoagulation. This is similar to the approach reported elsewhere.<sup>14 25</sup>

## Model validation

We tested the performance of the final model (QStroke) in the validation dataset in patients aged 25-84 years. We also compared QStroke with the Framingham stroke equation<sup>20</sup> for performance, restricting both samples to patients aged 35-74 years since this is the age range for which the Framingham equation was developed. We calculated the 10 year estimated risk of stroke or transient ischaemic attack for each patient in the validation dataset using multiple imputation to replace missing values as in the derivation dataset. We calculated the mean predicted and observed stroke risk at 10 years<sup>15</sup> and compared these by tenth of predicted risk for each score. The observed risk at 10 years was obtained using the 10 year Kaplan-Meier estimate. We calculated the receiver operating characteristics (ROC) statistic, D statistic (a measure of discrimination where higher values indicate better discrimination),<sup>26</sup> and an R<sup>2</sup> statistic (which is a measure of explained variation for survival data where higher values indicate more variation is explained).<sup>27</sup>

## Validation of QStroke, CHADS<sub>2</sub>, and CHA<sub>2</sub>DS<sub>2</sub>VASC in patients with atrial fibrillation

For patients with atrial fibrillation at baseline in the validation dataset, we calculated QStroke, CHADS<sub>2</sub>, and CHA<sub>2</sub>DS<sub>2</sub>VASC scores. We calculated Harrell's C statistic as a measure of discrimination in the subset of atrial fibrillation patients in the validation cohort, as this takes account of the censored nature of the data, unlike the ROC statistic. The C statistic was not calculated in the full validation cohort as the sample was too large and the test would not run. We also calculated the D statistic, and R<sup>2</sup> value using the numeric value of each score (QStroke, CHADS<sub>2</sub>, and CHA<sub>2</sub>DS<sub>2</sub>VASC).

We defined a high CHADS<sub>2</sub> or CHA<sub>2</sub>DS<sub>2</sub>VASC score as being  $\geq 1$ , since this is the cut-off value used to initiate anticoagulants. Since there is no currently accepted threshold for classifying high risk of stroke based on an absolute risk estimate, we examined the distribution of predicted risk values for QStroke and calculated a series of centile values which would identify similar numbers of patients to those identified using CHADS<sub>2</sub> or CHA<sub>2</sub>DS<sub>2</sub>VASC. We calculated the numbers and percentages of patients who would be reclassified using CHADS<sub>2</sub> or CHA<sub>2</sub>DS<sub>2</sub>VASC compared with QStroke (using the centile threshold values identified above). We calculated the observed risk of stroke or transient ischaemic attack at 10 years for each group of reclassified patients using Kaplan Meier estimates. There were at least 100 events per variable considered in the prediction modelling for the outcome in the derivation cohort.<sup>28</sup> Analyses were conducted using STATA (version 12).

## Results

### Practices and patients

Overall, 676 practices in England and Wales met our inclusion criteria and had been using their current computer system for at least one year. Of these, 451 were randomly assigned to the derivation dataset and 225 to the validation dataset. We identified 3 746 065 patients aged 25-84 years in the derivation cohort. Of these, 126 620 (3.4%) had missing Townsend scores, 47 425 (1.3%) had prior stroke or transient ischaemic attack, and 22 542 (0.6%) were prescribed oral anticoagulation at baseline, leaving 3 549 478 eligible patients. We identified 2 031 993 patients aged 25-84 years in the validation cohort. Of these 98 045 (4.8%) had missing Townsend scores, 24 463 (1.2%) had a prior diagnosis of stroke or transient ischaemic

attack, and 12 317 (0.6%) were prescribed oral anticoagulation at baseline, leaving 1 897 168 eligible patients.

Table 2<sup>11</sup> compares the characteristics of eligible patients in the derivation and validation cohorts. Although this validation cohort was drawn from an independent group of practices, the baseline characteristics were similar to those for the derivation cohort. For example, 50.6% of patients in the derivation cohort had ethnicity recorded compared with 50.8% in the validation cohort.

Of the 3 549 478 patients in the derivation cohort, 1 175 805 (33.1%) had at least 10 years of follow-up. Of the 1 897 168 patients in the validation cohort, 592 973 (31.3%) had at least 10 years of follow-up. The median follow-up was 7.0 years for the derivation cohort and 6.7 years for the validation cohort.

## Incidence of stroke

Table 3<sup>11</sup> shows the numbers of cases and incidence rates of stroke by age and sex in both cohorts and in the subset of patients with atrial fibrillation at baseline. Overall in the derivation cohort, we identified 77 578 incident strokes or transient ischaemic attacks arising from 24.8 million person years of observation. In the validation cohort we identified 38 404 incident cases of stroke arising from 12.7 million person years of observation. The incidence of stroke was similar in both men and women and in both the derivation and validation cohorts. As expected, the incidence rates were higher in the group of patients with atrial fibrillation.

## Model development

Table 4<sup>11</sup> shows the results of the Cox regression analysis for the final QStroke model. Details of the fractional polynomial terms for age and body mass index are shown in footnote of the table. The final model included interactions between age and the following variables in men and women: body mass index, systolic blood pressure, Townsend score, family history of coronary heart disease, coronary heart disease, congestive cardiac failure, treated hypertension, atrial fibrillation, type 1 diabetes, type 2 diabetes, valvular heart disease, and smoking status. There was also an interaction between age and atrial fibrillation in women but not men. The interactions with age indicated higher hazard ratios for these risk factors among younger patients compared with older patients, as with QRISK2.<sup>5</sup> Increasing material deprivation (as measured by the Townsend score) was associated with increasing stroke risk. There was a “dose-response” relationship for smoking, with heavy smokers having higher risks than moderate smokers, light smokers, or former smokers. Women in the Pakistani and Bangladeshi groups had significantly increased risks of stroke compared with women who were white or who didn’t have ethnicity recorded. Chinese men and men in the “other ethnic groups” has significantly lower risks of stroke compared with men who were white or who didn’t have ethnicity recorded. All the other factors in the table were significantly associated with increased stroke risk in men and women.

Of the 15 371 patients with atrial fibrillation at baseline in the derivation cohort, 3195 (20.8%) were subsequently prescribed anticoagulation during follow-up. Of the 7689 patients with atrial fibrillation in the validation cohort, 1640 (21.3%) were subsequently prescribed anticoagulation during follow-up. The results of the additional model in which patients with atrial fibrillation prescribed anticoagulation during follow-up were censored when they started treatment showed very similar hazard ratios to the main models presented here (results available from the authors).

## Calibration and discrimination of QStroke and Framingham stroke equations in the validation cohort

In the full validation cohort of people aged 25–84 years the QStroke algorithm explained 57% of the variation in women and 55% in men (table 5<sup>11</sup>). The D statistic was 2.4 in women and 2.3 in men. Table 5 also shows the corresponding results for the Framingham stroke equation in patients aged 35–74 years, with the comparison figures for QStroke in the same age range: the ROC values, R<sup>2</sup>, and D statistic values for QStroke were higher than those for Framingham. All the measures of performance were higher for women than men for both QStroke and Framingham.

Figure 1<sup>11</sup> compares the predicted and observed risks of stroke or transient ischaemic attack at 10 years using QStroke across each tenth of predicted risk (1 representing the lowest risk and 10 the highest risk) and demonstrates that the model is generally well calibrated for all patients free of stroke or transient ischaemic attack at baseline. The corresponding results for Framingham indicate a degree of under-prediction (fig 1<sup>11</sup>).

## Performance of QStroke, CHADS<sub>2</sub>, and CHA<sub>2</sub>DS<sub>2</sub>VASC in atrial fibrillation

Figure 2<sup>11</sup> compares the predicted and observed risks of stroke or transient ischaemic attack at 10 years in the subset of patients in the validation cohort with atrial fibrillation at baseline and shows that the model is well calibrated in men but there is a degree of over-prediction in women at higher levels of predicted risk.

Table 5<sup>11</sup> shows the validation statistics for QStroke, CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASC scores for men and women in the subset of 7689 patients in the validation cohort with atrial fibrillation at baseline. Of these, 890 had a stroke or transient ischaemic attack during follow-up. The point estimates in all measures of calibration and discrimination were higher in QStroke than CHA<sub>2</sub>DS<sub>2</sub>VASC and CHADS<sub>2</sub>, although the 95% confidence intervals were wide. For example, the C statistic in men was 0.71 for QStroke, 0.67 for CHA<sub>2</sub>DS<sub>2</sub>VASC, and 0.63 for CHADS<sub>2</sub>. The R<sup>2</sup> statistic in men was 24.1% for QStroke, 18.3% for CHA<sub>2</sub>DS<sub>2</sub>VASC, and 13.5% for CHADS<sub>2</sub>. The D statistic in men was 1.15 for QStroke, 0.97 for CHA<sub>2</sub>DS<sub>2</sub>VASC, and 0.81 for CHADS<sub>2</sub>. The validation statistics for men tended to be higher than for women with atrial fibrillation for each of the three scores.

Table 6<sup>11</sup> shows the performance statistics for QStroke, CHADS<sub>2</sub>, and CHA<sub>2</sub>DS<sub>2</sub>VASC for the patients with atrial fibrillation at baseline in the validation cohort. For CHADS<sub>2</sub>, 63% of patients had a score of  $\geq 1$  so were classified as high risk. The sensitivity was 77% and the observed 10 year risk was 22.8%. We identified the top 63% of men and women with the highest QStroke scores in order to assemble a group of comparable size to those classified at high risk using CHADS<sub>2</sub>. This was equivalent to a 10 year risk threshold of 15%. Using this definition, QStroke had a sensitivity of 83%, and the observed 10 year risk in this group was 24.4%.

For CHA<sub>2</sub>DS<sub>2</sub>VASC, 85% of men and women had a score of  $\geq 1$  so were classified as high risk. The sensitivity at this threshold was 97%, and the observed 10 year risk was 20.2%. Similarly, we identified the top 85% of patients with the highest QStroke scores in order to assemble a group of comparable size to those classified at high risk using CHA<sub>2</sub>DS<sub>2</sub>VASC. This was equivalent to a 10 year risk threshold of 5.1%. Using this definition, QStroke had a sensitivity of 98%, and a 10 year observed risk of 20.5%.

Table 7 shows the reclassification statistics for a high CHADS<sub>2</sub> score compared with a high QStroke 10 year risk score based on the top 63% of patients at highest risk. Of the 7689 patients with atrial fibrillation, 2195 (29%) were classified as low risk on both CHADS<sub>2</sub> and QStroke. The observed 10 year risk of stroke in this group was 8%. There were 4187 (55%) patients who were high risk on both QStroke and CHADS<sub>2</sub>. The observed 10 year risk of stroke in this group was 25%. There were 657 (9%) patients who were high risk on QStroke and low risk on CHADS<sub>2</sub>; these patients had an observed 10 year risk of stroke of 19%. There were 650 patients (9%) who were low risk on QStroke but high risk on CHADS<sub>2</sub>; these patients had a 10 year absolute risk of 8%.

Table 7 also shows the reclassification statistics for a high CHA<sub>2</sub>DS<sub>2</sub>VASc score compared with a high QStroke 10 year risk score based on the top 85% of patients at highest risk. Overall 4% of patients would be reclassified from low to high risk using QStroke compared with CHA<sub>2</sub>DS<sub>2</sub>VASc—the observed risk in these patients was 8%. Similarly 4% of patients would be reclassified from high to low risk using QStroke compared with CHA<sub>2</sub>DS<sub>2</sub>VASc, and these had an observed risk of 3%.

## Discussion

### Summary of key findings

We have developed and validated QStroke, which is a new algorithm to identify patients at high risk of ischaemic stroke based on contemporaneous primary care data from the UK. Although QStroke has been designed to be used in all patients without a history of stroke or transient ischaemic attack, we envisage that its primary use will be in the subset of patients with atrial fibrillation for whom anticoagulation is considered.

QStroke incorporates established risk factors for stroke or transient ischaemic attack, many of which are absent from existing stroke risk assessment tools. QStroke includes age, sex, deprivation, ethnicity, body mass index, systolic blood pressure, total cholesterol:HDL cholesterol ratio, smoking status (five levels), diabetes type, congestive cardiac failure, coronary heart disease, rheumatoid arthritis, chronic kidney disease, treated hypertension, valvular heart disease, and family history of premature coronary heart disease.

### Comparison with existing risk prediction scores

We tested the performance of QStroke in a separate cohort of patients without stroke or transient ischaemic attack and demonstrated good levels of discrimination and calibration and improved performance compared with the Framingham stroke risk score.

We also tested QStroke in the subset of patients with atrial fibrillation, for whom anticoagulation might be indicated. We compared the performance of QStroke with both CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASc in patients without a prior stroke to ensure a fair comparison between the scores. We found some indication of improved performance on all measures of discrimination, although confidence intervals were wide. The comparison between QStroke and CHADS<sub>2</sub> is important since the use of CHADS<sub>2</sub> is currently incentivised as an indicator in the primary care Quality and Outcomes Framework and is used to determine which patients require anticoagulation. We also demonstrated some evidence of improved performance of QStroke compared with the newer CHA<sub>2</sub>DS<sub>2</sub>VASc, although this was less marked. None the less, we think the difference between performance of

QStroke and CHA<sub>2</sub>DS<sub>2</sub>VASc could be important for those patients who are reclassified with QStroke and for whom advice on treatment with anticoagulation might change. For example, patients at high predicted risk on QStroke but classified as low risk on CHA<sub>2</sub>DS<sub>2</sub>VASc might require anticoagulation. Conversely, the patients classified as high risk with CHA<sub>2</sub>DS<sub>2</sub>VASc but low predicted risk with QStroke might be able to avoid unnecessary anticoagulation.

We have not provided definite comment on what threshold of absolute risk should be used for intervention, as that would include cost effectiveness analyses, which are outside the scope of this study. Ideally the review of a risk score is best judged around a risk threshold. This can be appropriate for other scores such as QRISK2 determining whether to intervene with primary prevention of cardiovascular disease. When determining whether to intervene with primary prevention of cardiovascular disease, the current 10 year risk threshold is 20%, making comparisons of different risk scores such as QRISK2 around this threshold appropriate. Stroke prevention in the population with atrial fibrillation has not yet reached this level of sophistication. The currently accepted risk scores, such as those of CHADS<sub>2</sub>, have not described their outputs in terms of absolute risk of stroke, and, as such, there is no consensus regarding a risk threshold. In contrast, QStroke calculates the absolute risk of stroke and so, unlike CHADS<sub>2</sub>, is able to inform future debate around what threshold is appropriate to intervene with oral anticoagulation. Choice of threshold is a complex area dependent on many variables relating to clinical outcomes and service costs, and to do justice to the complexity, we consider it should be the subject of a separate paper. We have, however, provided analyses using a range of thresholds of risk which can be used to help inform future analyses and guidelines.

The results of our validation statistics for CHA<sub>2</sub>DS<sub>2</sub>VASc and CHADS<sub>2</sub> in patients with atrial fibrillation are broadly similar to those reported using another UK GP database<sup>25</sup> and a Danish registry cohort.<sup>14</sup> Both studies showed improved performance of CHA<sub>2</sub>DS<sub>2</sub>VASc compared with CHADS<sub>2</sub>. The Danish study additionally showed that CHA<sub>2</sub>DS<sub>2</sub>VASc was better for identifying those at low and intermediate risk.<sup>14</sup> Our results, however, are not directly comparable with those of the Danish study as that study included venous thromboembolism in the definition of the outcome, whereas our study included only stroke and transient ischaemic attack.<sup>14</sup> QStroke, CHA<sub>2</sub>DS<sub>2</sub>VASc, and CHADS<sub>2</sub> tended to perform better in men with atrial fibrillation compared with women with atrial fibrillation, which deserves further study.

### Implications for clinical practice

Whilst the new QStroke algorithm is more complex than CHA<sub>2</sub>DS<sub>2</sub>VASc or CHADS<sub>2</sub>, it has several advantages. It includes weighting for ethnicity and deprivation, which should help avoid widening health inequalities. The algorithm uses routinely collected data, which means it can be easily and regularly updated to reflect changes in populations, improvements in data quality, advances in knowledge, and evolving guidelines. The algorithms can also be implemented in primary care since the data are already present in the clinical computer systems. QStroke will work both in populations with atrial fibrillation and those without atrial fibrillation—though the immediate clinical use might be for risk stratification among patients with atrial fibrillation, QStroke can still inform other patients of their specific risk of stroke or transient ischaemic attack as part of their general cardiovascular risk assessment.

QStroke has also been designed to be integrated into UK general practice clinical computer systems, where the risk factors are already recorded and used to calculate closely related scores such as QRISK2. Much of the apparent complexity relating to additional variables and interactions can be incorporated into the software using data already entered into each patient's electronic health record. There are only three variables in QStroke (congestive cardiac failure, coronary heart disease, and valvular heart disease) that are not in QRISK2. Where possible we used the definitions from the Quality and Outcomes Framework, which should simplify its implementation. QRISK2 is integrated into all four UK GP clinical computer systems, and QStroke can be implemented in a similar way. For example, clinicians can use structured templates within the consultation to calculate a patient's risk and use the information to inform treatment decisions. It can also be used in "batch processing" mode to calculate an estimated risk for all eligible patients registered with a practice so that patients with the highest risk can be recalled. Additionally, QStroke could easily be integrated in the GRASP-AF tool (Guidance on Risk Assessment and Stroke Prevention in Atrial Fibrillation), which is a primary care database interrogation tool designed to help identify possible candidates for anticoagulation from practice lists.<sup>29</sup>

Another advantage of QStroke compared with either CHA<sub>2</sub>DS<sub>2</sub>VASc or CHADS<sub>2</sub> is that it gives an absolute measure of stroke risk which can more easily be explained to a patients (for example, "Of 100 people like you, X are likely to have a stroke or transient ischaemic attack within the next 10 years"), rather than a simple integer that has no direct interpretation of absolute stroke risk. Should a tool be developed that quantifies absolute risk of bleeding with anticoagulation, it will be possible to do a more direct assessment of risk of stroke in patients compared with potential risk and benefits of anticoagulation, thus providing better information for patients to make an informed choice. This is important since anticoagulation treatment is usually life long, and the risk of bleeding increases with increasing age.

QStroke has not been designed be used in patients with atrial fibrillation who have had a previous stroke, since all such patients should be prescribed anticoagulation and an estimation of risk will not affect the clinical decision. To ensure a fair comparison, we compared the performance of QStroke against CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASc only in patients with atrial fibrillation who were free from stroke. However, removing the patients already receiving treatment may result in the higher risk patients being removed from the cohort, which might then result in an underestimation of risk in patients with atrial fibrillation overall.

## Methodological considerations

The methods to derive and validate this model are the same as those used for the original development of QRISK2 and a range of other risk prediction tools. The strengths and limitations of the approach have already been discussed in detail,<sup>4 30-34</sup> including information on multiple imputation of missing data. In summary, key strengths include size, duration of follow-up, representativeness, and lack of selection, recall, and respondent bias. UK general practices have good levels of accuracy and completeness in recording clinical diagnoses and prescribed drugs.<sup>35 36</sup> We think our study has good face validity since it has been conducted in the setting where most patients in the UK are assessed, treated, and followed up. Limitations include lack of formally adjudicated outcomes, information bias, and potential for bias due to missing data. Our database has linked cause of death from the UK Office of National Statistics, and

our study is therefore likely to have picked up most cases of stroke or transient ischaemic attack, thereby minimising ascertainment bias. Patients who die of stroke in hospital will have stroke or transient ischaemic attack recorded on their death certificate and therefore will be included on the linked cause of death data. Other patients who have stroke or transient ischaemic attack diagnosed in hospital who do not die will have the information recorded in hospital discharge letters which are sent to the patients' general practice and then entered into each patient's electronic record. We excluded people without a valid deprivation score since this group may represent a more transient population where follow-up for stroke could be unreliable or unrepresentative. Their deprivation scores are unlikely to be missing at random so we did not think it would be appropriate to impute them.

The present validation has been done on a completely separate set of practices and individuals to those which were used to develop the score, although the practices all use the same clinical computer system (EMIS, the computer system used by 55% of UK general practices). An independent validation study would be a more stringent test and should be done, but when such independent studies have examined other risk algorithms,<sup>6 8 31 33</sup> they have demonstrated similar performance compared with the validation in the QResearch database.<sup>5 30 32</sup>

This QStroke model has been developed using data from England and Wales and includes UK derived ethnicities and a postcode-based deprivation score. It is therefore not immediately applicable for clinical use in international settings without some modification of the UK-specific risk factors and validation in the setting in which it is intended to be used.

## Conclusions

We have developed and validated a new algorithm to predict risk of stroke. QStroke shows some improvement over current risk scoring methods, CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASc, for patients with atrial fibrillation for whom anticoagulation may be required. QStroke also provides an accurate measure of absolute stroke risk in the general population of patients free of stroke or transient ischaemic attack, as shown by its performance in a separate validation cohort. Further research is needed to evaluate the clinical outcomes and cost effectiveness of using these algorithms in primary care.

We acknowledge the contribution of EMIS practices that contribute to QResearch and the University of Nottingham and EMIS for expertise in establishing, developing, and supporting the database.

**Contributors:** JHC initiated the study; undertook the literature review, data extraction, data manipulation, and primary data analysis; and wrote the first draft of the paper. CC contributed to the design, analysis, interpretation, and drafting of the paper. PB contributed to the development of core ideas, the analysis plan, the interpretation of the results, and the drafting of the paper.

**Funding:** None.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at [www.icmje.org/coi\\_disclosure.pdf](http://www.icmje.org/coi_disclosure.pdf) (available on request from the corresponding author) and declare: JHC is professor of clinical epidemiology at the University of Nottingham and co-director of QResearch, a not-for-profit organisation which is a joint partnership between the University of Nottingham and EMIS (commercial supplier of IT for 60% of general practices in the UK). JHC is also director of ClinRisk, which produces open and closed source software to ensure reliable and updatable implementation of clinical risk algorithms within clinical computer systems. CC is associate professor of Medical Statistics at the University of Nottingham and a consultant statistician for ClinRisk.

**What is already known on this topic**

Methods to identify patients at high or low risk of stroke are needed to identify patients for whom interventions may be required, especially those with atrial fibrillation for whom anticoagulation might be needed

Current methods for risk scoring, such as CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASc, are not based on a statistical model, do not include many established risk factors, nor provide absolute risk estimates of stroke

**What this study adds**

We have developed a new algorithm to quantify absolute risk of primary stroke which includes established risk factors and which is designed to work with the QRISK2 cardiovascular disease algorithm

QStroke provides a valid measure of absolute stroke risk in the general population of patients free of stroke or transient ischaemic attack as shown by its performance in a separate validation cohort

QStroke shows some improvement on current risk scoring methods, CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>VASc, for the subset of patients with atrial fibrillation for whom anticoagulation may be required

Further research is needed to evaluate the clinical outcomes and cost effectiveness of using these algorithms in primary care

This work and any views expressed within it are solely those of the authors and not of any affiliated bodies or organisations. There are no other relationships or activities that could appear to have influenced the submitted work.

**Approvals:** The project was approved in accordance with the QResearch agreement with Trent Multi-Centre Research Ethics Committee.

**Data sharing:** The patient level data from the QResearch are specifically licensed according to its governance framework. See [www.qresearch.org](http://www.qresearch.org) for further details. The QStroke algorithm will be published as open source software under the GNU Lesser Public License.

- 1 British Heart Foundation. Coronary heart disease statistics. Secondary coronary heart disease statistics. 2007. [www.heartstats.org/homepage.asp](http://www.heartstats.org/homepage.asp).
- 2 Vascular Programme. *Putting prevention first. Vascular checks: risk assessment and management*. Department of Health, 2008;15.
- 3 National Collaborating Centre for Primary Care. Section 4.3 of the Guideline on Cardiovascular Risk Assessment: the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. NICE, 2008;43.
- 4 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. 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.
- 5 Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008;336:1475-82.
- 6 Collins GS, Altman A. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* 2012;344:e4181.
- 7 Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009;339:b2584.
- 8 Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010;340:c2442.
- 9 Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Ann Intern Med* 2007;146:857-67.
- 10 Health and Social Care Information Centre. Quality and Outcomes Framework. [www.ic.nhs.uk/services/qof](http://www.ic.nhs.uk/services/qof).
- 11 Lip GY. Atrial fibrillation in 2011: stroke prevention in AF. *Nat Rev Cardiol* 2011;9:71-3.
- 12 Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: Results from the national registry of atrial fibrillation. *JAMA* 2001;285:2864-70.
- 13 Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* 2010;137:263-72.
- 14 Olesen JB, Lip GYH, Hansen ML, Hansen PR, Tolstrup JS, Lindhardsen J, et al. Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. *BMJ* 2011;342:d124.
- 15 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;335:136.
- 16 Hippisley-Cox J, Vinogradova Y, Coupland C, Pringle M. Comparison of key practice characteristics between general practices in England and Wales and general practices in the QRESEARCH data. Report to the Health and Social Care Information Centre. University of Nottingham, 2005.
- 17 Camm AJ, Lip GY, De Caterina R, Savelieva I, Atar D, Hohnloser SH, et al. 2012 focused update of the ESC Guidelines for the management of atrial fibrillation: an update of the

- 18 2010 ESC Guidelines for the management of atrial fibrillation—developed with the special contribution of the European Heart Rhythm Association. *Europace* 2012;14:1385-413.
- 19 Wang TJ, Larson MG, Levy D, Vasan RS, Leip EP, Wolf PA, et al. Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality: the Framingham Heart Study. *Circulation* 2003;107:2920-5.
- 20 Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham study. *Stroke* 1991;22:312-8.
- 21 Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991;121:293-8.
- 22 Maradit-Kremers H, Crowson CS, Nicola PJ, Ballman KV, Roger VL, Jacobsen SJ, et al. Increased unrecognized coronary heart disease and sudden deaths in rheumatoid arthritis: a population-based cohort study. *Arthritis Rheum* 2005;52:402-11.
- 23 Sarnak MJ, Levey AS, Schoolwerth AC, Coresh J, Culleton B, Hamm LL, et al. Kidney disease as a risk factor for development of cardiovascular disease: a statement from the American Heart Association Councils on Kidney in Cardiovascular Disease, High Blood Pressure Research, Clinical Cardiology, and Epidemiology and Prevention. *Circulation* 2003;108:2154-69.
- 24 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964-74.
- 25 Royston P. Multiple imputation of missing values. *Stata J* 2004;4:227-41.
- 26 Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723-48.
- 27 Royston P. Explained variation for survival models. *Stata J* 2006;6:1-14.
- 28 Steyerberg E. *Clinical prediction models*. Springer, 2009.
- 29 NHS Improvement. Guidance on risk assessment and stroke prevention for atrial fibrillation (GRASP-AF). Secondary guidance on risk assessment and stroke prevention for atrial fibrillation (GRASP-AF). 2012. [www.improvement.nhs.uk/graspaf/GRASPRResources.html](http://www.improvement.nhs.uk/graspaf/GRASPRResources.html).
- 30 Hippisley-Cox J, Coupland C. Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *BMJ* 2009;339:b4229.
- 31 Collins GS, Mallett S, Altman DG. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. *BMJ* 2011;342:d3651.
- 32 Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009;338:b880.
- 33 Collins GS, Altman DG. External validation of the QDScore for predicting the 10-year risk of developing type 2 diabetes. *Diabet Med* 2011;28:599-607.
- 34 Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithm (QThrombosis) to estimate future risk of venous thromboembolism: prospective cohort study. *BMJ* 2011;343:d4656.
- 35 Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991;302:766-8.
- 36 Majeed A. Sources, uses, strengths and limitations of data collected in primary care in England. *Health Stat Q* 2004;(21):5-14.

**Accepted:** 22 March 2013

**Cite this as:** *BMJ* 2013;346:f2573

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>.

## Tables

**Table 1 | Variables included in risk algorithms QStroke (2013), Framingham (1991), CHA<sub>2</sub>DS<sub>2</sub>VASc, and CHADS<sub>2</sub>**

	QStroke (2013) (range values 0-99%)	Framingham (1991) (range values 0-99%*)	CHA <sub>2</sub> DS <sub>2</sub> VASc (range values 0-9)	CHADS <sub>2</sub> (range values 0-6)
Age (years)	Continuous (range 25-84 years)	Continuous (range 35-74 years)	Categorical (1 point for 65-74, Binary (1 point if >75 years) 2 points for ≥75 years)	
Sex	Yes, separate models in men and women	Yes	Yes, 1 point if female	No
Treated hypertension	Yes, binary	No	Yes, 1 point	Yes, 1 point
Type 1 diabetes	Yes, binary	Yes, types 1 and 2 combined	Yes, 1 point for type 1 or 2	Yes, 1 point for type 1 or 2
Type 2 diabetes	Yes, binary			
Atrial fibrillation	Yes, binary	No	Selection criteria	Selection criteria
Congestive cardiac failure	Yes, binary	No	Yes, 1 point	Yes, 1 point
Coronary heart disease†	Yes, binary	No	Yes, 1 point	No
Prior stroke or TIA	No	No	Yes, 2 points	Yes, 2 points
Self assigned ethnicity	Yes, 9 categories	No	No	No
Deprivation score	Yes, continuous	No	No	No
Smoking status	Yes, 5 categories	Binary	No	No
Systolic blood pressure	Yes, continuous	Yes, continuous	No	No
Cholesterol:HDL ratio	Yes, continuous	Yes, continuous	No	No
Body mass index	Yes, continuous	No	No	No
Family history of coronary disease	Yes, binary	No	No	No
Rheumatoid arthritis	Yes, binary	No	No	No
Chronic renal disease	Yes, binary	No	No	No
Valvular heart disease	Yes, binary	No	No	No

TIA=transient ischaemic attack. HDL=high density lipoprotein cholesterol.

\*Framingham also includes left ventricular hypertrophy.

†CHA<sub>2</sub>DS<sub>2</sub>VASc includes prior vascular disease defined as myocardial infarction, peripheral arterial disease, or aortic plaque.<sup>14</sup>

**Table 2** Baseline characteristics of the derivation and validation cohorts (1998–2012) used in development and validation of QStroke risk algorithm. Values are numbers (percentages) unless stated otherwise

Characteristic	Derivation cohort (n=3 549 478)	Validation cohort (n=1 897 168)
Women	1 801 370 (50.8)	962 083 (50.7)
Men	1 748 108 (49.2)	935 085 (49.3)
Mean (SD) age (years)	45.0 (15.4)	44.9 (15.3)
Mean (SD) Townsend deprivation score	-0.1 (3.5)	0.2 (3.6)
Ethnicity:		
Recorded	1 797 073 (50.6)	962 961 (50.8)
White or not recorded	3 333 695 (93.9)	1 760 458 (92.8)
Indian	43 175 (1.2)	27 096 (1.4)
Pakistani	21 101 (0.6)	14 131 (0.7)
Bangladeshi	13 878 (0.4)	9 404 (0.5)
Other Asian	26 435 (0.7)	14 743 (0.8)
Caribbean	17 488 (0.5)	13 296 (0.7)
Black African	36 952 (1.0)	24 002 (1.3)
Chinese	13 862 (0.4)	7 296 (0.4)
Other ethnic group	42 892 (1.2)	26 742 (1.4)
Medical factors:		
Atrial fibrillation	15 371 (0.4)	7 689 (0.4)
Family history of CHD at age <60 years	424 745 (12.0)	195 482 (10.3)
Type 1 diabetes	11 636 (0.3)	6 058 (0.3)
Type 2 diabetes	79 567 (2.2)	44 267 (2.3)
Rheumatoid arthritis	20 580 (0.6)	10 777 (0.6)
Chronic renal disease	6 536 (0.2)	3 749 (0.2)
Treated hypertension	236 900 (6.7)	124 881 (6.6)
Coronary heart disease	99 561 (2.8)	53 328 (2.8)
Congestive cardiac failure	16 294 (0.5)	8 390 (0.4)
Valvular heart disease	13 510 (0.4)	7 186 (0.4)
Prescribed aspirin	129 648 (3.7)	71 392 (3.8)
Body mass index recorded	2 804 840 (79.0)	1 481 616 (78.1)
Mean (SD) body mass index	26.0 (4.6)	26.0 (4.6)
Systolic blood pressure recorded	3 113 490 (87.7)	1 646 890 (86.8)
Mean (SD) systolic blood pressure (mm Hg)	129.9 (19.9)	129.7 (19.7)
Total cholesterol:HDL cholesterol ratio recorded	1 160 290 (32.7)	642 419 (33.9)
Mean (SD) total cholesterol:HDL cholesterol ratio	4.1 (1.3)	4.1 (1.3)
Smoking not recorded	298 902 (8.4)	179 523 (9.5)
Smoking recorded	3 250 576 (91.6)	1 717 645 (90.5)
Non-smoker	1 789 104 (50.4)	955 968 (50.4)
Former smoker	595 287 (16.8)	300 905 (15.9)
Light smoker	246 174 (6.9)	129 974 (6.9)
Moderate smoker	293 150 (8.3)	159 204 (8.4)
Heavy smoker	203 303 (5.7)	109 521 (5.8)
Current smoker, amount not recorded	123 558 (3.5)	62 073 (3.3)

CHD=coronary heart disease. HDL=high density lipoprotein.

**Table 3** Incidence rates of ischaemic stroke or transient ischaemic attack for men and women in the derivation and validation cohorts.

Rates are per 100 000 person years.

Age (years)	Women: derivation cohort			Women: validation cohort			Men: derivation cohort			Men: validation cohort		
	No of cases	Person years	Incidence rate	No of cases	Person years	Incidence rate	No of cases	Person years	Incidence rate	No of cases	Person years	Incidence rate
Total population:	39 504	12 623 597	3.1	19 447	6 480 823	3.0	38 074	12 150 749	3.1	18 957	6 240 093	3.0
25-34	616	3 145 897	0.2	295	1 607 038	0.2	580	3 119 427	0.2	279	1 596 364	0.2
35-44	1 573	2 943 382	0.5	761	1 542 845	0.5	1 998	3 104 291	0.6	999	1 607 242	0.6
45-54	3 464	2 500 376	1.4	1 773	1 263 734	1.4	5 098	2 553 386	2.0	2 596	1 295 034	2.0
55-64	6 871	1 897 844	3.6	3 554	979 076	3.6	9 354	1 814 064	5.2	4 584	936 382	4.9
65-74	13 028	1 394 488	9.3	6 418	715 614	9.0	12 513	1 129 222	11.1	6 323	583 441	10.8
≥75	13 952	741 610	18.8	6 646	372 516	17.8	8 531	430 359	19.8	4 176	221 631	18.8
Subset with AF:	987	38 668	25.5	416	19 192	21.7	905	53 264	17.0	474	26 485	17.9
25-34	0	362	0.0	1	227	4.4	1	1 149	0.9	0	819	0
35-44	4	1 076	3.7	1	354	2.8	5	3 746	1.3	2	1 846	1.1
45-54	11	1 915	5.7	4	1 200	3.3	30	6 350	4.7	12	3 302	3.6
55-64	55	5 390	10.2	28	2 911	9.6	128	11 575	11.1	62	5 721	10.8
65-74	280	12 123	23.1	106	6 297	16.8	338	17 369	19.5	163	8 113	20.1
≥75	637	17 803	35.8	276	8 203	33.7	403	13 076	30.8	235	6 684	35.2

AF=atrial fibrillation.

**Table 4 | Adjusted hazard ratios (95% CI) for stroke or transient ischaemic attack for the final QStroke model in the derivation cohort. Hazard ratios are adjusted for fractional polynomial terms for age and body mass index. Final model included age interaction terms**

Variable	Adjusted hazard ratio (95% CI)*	
	Women	Men
Systolic blood pressure (10 unit increase)	1.12 (1.10 to 1.13)	1.15 (1.14 to 1.16)
Total cholesterol:HDL cholesterol ratio (1 unit increase)	1.08 (1.07 to 1.09)	1.08 (1.07 to 1.09)
Townsend deprivation score (5 unit increase)	1.33 (1.28 to 1.37)	1.23 (1.19 to 1.27)
Smoking status:		
Non-smoker	1.00	1.00
Former smoker	1.16 (1.08 to 1.25)	1.17 (1.1 to 1.24)
Light smoker	1.57 (1.44 to 1.71)	1.70 (1.58 to 1.82)
Moderate smoker	1.85 (1.73 to 1.98)	1.70 (1.59 to 1.83)
Heavy smoker	2.25 (2.10 to 2.42)	2.06 (1.93 to 2.21)
Ethnicity:		
White or not recorded	1.00	1.00
Indian	0.97 (0.84 to 1.12)	0.93 (0.82 to 1.05)
Pakistani	1.56 (1.30 to 1.88)	1.06 (0.88 to 1.27)
Bangladeshi	1.31 (1.02 to 1.67)	1.08 (0.87 to 1.33)
Other Asian	0.77 (0.60 to 1.01)	0.98 (0.78 to 1.22)
Caribbean	0.95 (0.82 to 1.11)	0.98 (0.84 to 1.14)
Black African	0.76 (0.58 to 0.99)	0.84 (0.66 to 1.08)
Chinese	0.51 (0.32 to 0.80)	0.66 (0.44 to 1.00)
Other ethnic group	0.85 (0.71 to 1.02)	0.74 (0.61 to 0.89)
Medical and family history:		
Family history of coronary heart disease†	1.33 (1.26 to 1.41)	1.29 (1.22 to 1.37)
Atrial fibrillation†	3.08 (1.89 to 5.00)	1.60 (1.49 to 1.71)
Coronary heart disease†	3.12 (2.60 to 3.74)	2.56 (2.27 to 2.90)
Congestive cardiac failure†	2.72 (1.73 to 4.28)	2.42 (1.81 to 3.25)
Type 1 diabetes†	3.64 (2.91 to 4.56)	3.62 (2.93 to 4.46)
Type 2 diabetes†	2.17 (1.85 to 2.54)	1.97 (1.72 to 2.25)
Treated hypertension†	1.82 (1.66 to 2.00)	1.84 (1.71 to 1.99)
Rheumatoid arthritis†	1.34 (1.25 to 1.43)	1.24 (1.12 to 1.38)
Chronic renal disease†	1.47 (1.25 to 1.72)	1.39 (1.21 to 1.61)
Valvular heart disease†	2.42 (1.89 to 3.09)	2.40 (1.86 to 3.10)

\*Hazard ratios simultaneously adjusted for all the other variables shown in the table as well as fractional polynomial terms for age and body mass index.

†Compared with patients without the condition or medication at baseline.

Models also included fractional polynomial terms for age and body mass index (BMI). For women, fractional polynomial terms were  $(age/10)^2$  and  $(age/10)^3$ ;  $(BMI/10)^{-2}$  and  $(BMI/10)^{-2} \ln(BMI)$ . For men, fractional polynomial terms were  $(age/10)^{-1}$  and  $(age/10)^{-1} \ln(age)$ ;  $(BMI/10)^{-2}$  and  $(BMI/10)^{-2} \ln(BMI)$ .

The models for men and women also included interactions between the age terms and BMI, systolic blood pressure, Townsend score, family history of coronary heart disease, coronary heart disease, congestive cardiac failure, treated hypertension, atrial fibrillation (women only), valvular heart disease, type 1 diabetes, type 2 diabetes, and smoking. Hazard ratios for these variables in the table are evaluated at mean age in men and women.

**Table 5| Validation statistics for the QStroke prediction algorithm in the validation cohort**

Prediction algorithm and validation statistic	Mean (95% CI)	
	Women (n=962 083)	Men (n=935 085)
All patients free of stroke or TIA at baseline		
QStroke (25-84 years):		
R <sup>2</sup> (%)	57.3 (56.8 to 57.8)	55.1 (54.6 to 55.7)
D statistic	2.37 (2.35 to 2.40)	2.27 (2.24 to 2.30)
ROC statistic	0.877 (0.875 to 0.879)	0.866 (0.864 to 0.868)
QStroke (35-74 years):		
R <sup>2</sup> (%)	43.7 (42.9 to 44.5)	41.9 (41.1 to 42.7)
D statistic	1.80 (1.77 to 1.83)	1.74 (1.71 to 1.77)
ROC statistic	0.814 (0.810 to 0.818)	0.806 (0.802 to 0.809)
Framingham stroke equation (35-74 years):		
R <sup>2</sup> (%)	38.5 (37.7 to 39.4)	35.7 (34.8 to 36.5)
D statistic	1.62 (1.59 to 1.65)	1.52 (1.50 to 1.55)
ROC statistic	0.798 (0.794 to 0.802)	0.788 (0.784 to 0.791)
Patients with atrial fibrillation at baseline	(n=3180)	(n=4509)
QStroke:		
R <sup>2</sup> (%)	14.0 (9.2 to 18.7)	24.1 (19.3 to 28.9)
D statistic	0.82 (0.66 to 0.99)	1.15 (1.00 to 1.30)
Harrell's C statistic	0.65 (0.62 to 0.67)	0.71 (0.69 to 0.73)
CHA <sub>2</sub> DS <sub>2</sub> VASc:		
R <sup>2</sup> (%)	9.6 (5.5 to 13.8)	18.3 (13.7 to 22.8)
D statistic	0.67 (0.51 to 0.83)	0.97 (0.82 to 1.12)
Harrell's C statistic	0.62 (0.59 to 0.65)	0.67 (0.65 to 0.69)
CHADS <sub>2</sub> :		
R <sup>2</sup> (%)	9.1 (4.9 to 13.2)	13.5 (9.1 to 17.9)
D statistic	0.64 (0.49 to 0.81)	0.81 (0.66 to 0.96)
Harrell's C statistic	0.61 (0.59 to 0.64)	0.63 (0.61 to 0.66)

TIA=transient ischaemic attack.

D statistic is a measure of discrimination—higher values indicate better discrimination.

ROC statistic is a measure of discrimination—higher values indicate better discrimination.

Harrell's C statistic is an extension to the ROC statistic which takes account of censored data—higher values indicate better discrimination. It was not possible to calculate Harrell's C in the cohort of all patients without because of the large sample size.

**Table 6| Performance of CHADS<sub>2</sub>, CHA<sub>2</sub>DS<sub>2</sub>VASc, and QStroke risk algorithms for predicting stroke in 7689 patients with atrial fibrillation from the validation cohort**

	Cut off 10 year predicted risk (%)	No (%) classified as high risk	Sensitivity (%)	Specificity (%)*	Observed 10 year risk (%)
CHADS <sub>2</sub> score ≥1	N/A	4837 (62.9)	76.9	38.9	22.8
QStroke, top 63% at high risk	15.1	4844 (63.0)	82.5	39.5	24.4
CHA <sub>2</sub> DS <sub>2</sub> VASc score ≥1	N/A	6543 (85.1)	96.6	16.4	20.2
QStroke, top 85% at high risk	5.1	6535 (85.0)	97.9	16.7	20.5
QStroke, top 70% at high risk	12.0	5382 (70.0)	89.0	32.5	23.3
QStroke, top 80% at high risk	7.6	6151 (80.0)	95.8	22.1	21.3
QStroke, top 90% at high risk	2.7	6920 (90.0)	99.2	11.2	19.4

\*Specificity is only approximate since it does not take account of censored data, and 1608 (21%) of patients were censored before 10 years. So patients counted as not having a stroke in the specificity calculation may have had a stroke within 10 years but after they were censored.

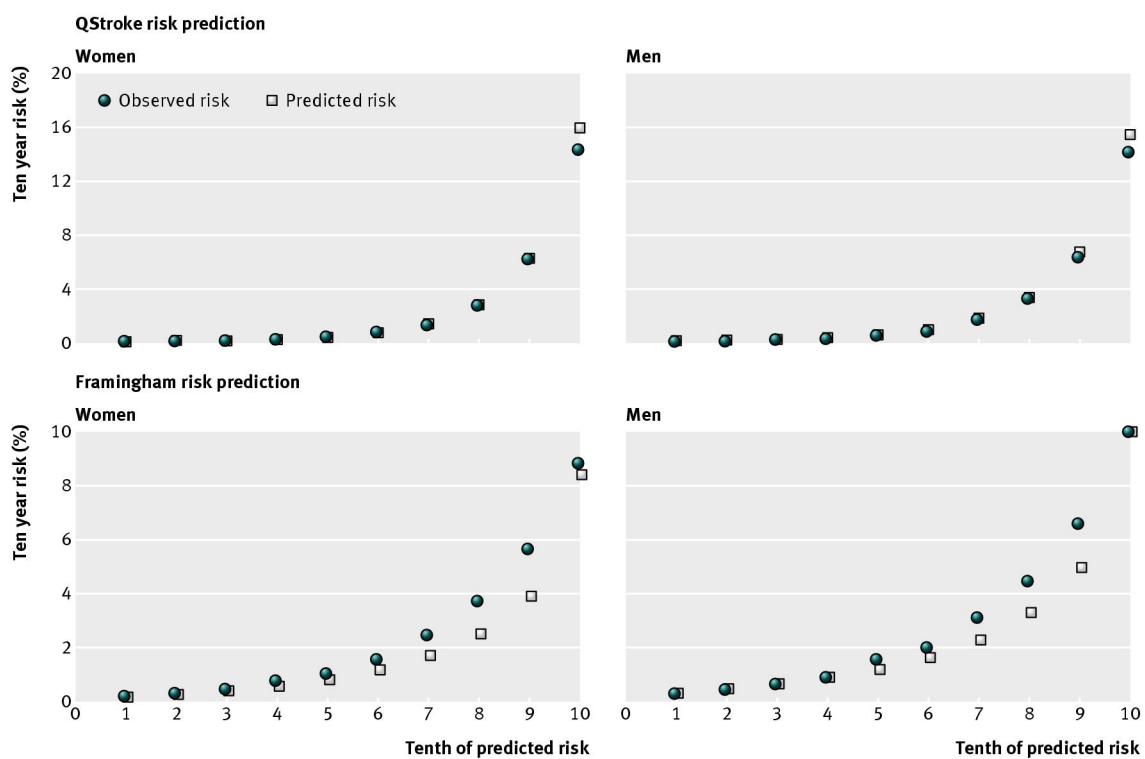
**Table 7** Reclassification of 7689 patients with atrial fibrillation in the validation cohort based on QStroke and high CHADS<sub>2</sub> score or high CHA<sub>2</sub>DS<sub>2</sub>VASc score

	No (%) of patients	10 year observed risk (95% CI)
<b>QStroke v CHADS<sub>2</sub>*</b>		
Low risk on QStroke* and CHADS <sub>2</sub>	2195 (28.6)	7.6 (6.2 to 9.2)
High risk on QStroke, low risk on CHADS <sub>2</sub>	657 (8.5)	19.3 (15.6 to 23.7)
Low risk on QStroke, high risk on CHADS <sub>2</sub>	650 (8.5)	8.0 (5.4 to 11.4)
High risk on QStroke and CHADS <sub>2</sub>	4187 (54.5)	25.4 (23.4 to 27.4)
<b>QStroke v CHA<sub>2</sub>DS<sub>2</sub>VASc†</b>		
Low risk on QStroke and CHA <sub>2</sub> DS <sub>2</sub> VASc	841 (10.9)	2.6 (1.5 to 4.6)
High risk on QStroke, low risk on CHA <sub>2</sub> DS <sub>2</sub> VASc	305 (4.0)	7.6 (4.6 to 12.5)
Low risk on QStroke, high risk on CHA <sub>2</sub> DS <sub>2</sub> VASc	313 (4.1)	2.8 (1.0 to 7.4)
High risk on QStroke and CHA <sub>2</sub> DS <sub>2</sub> VASc	6230 (81.0)	21.2 (19.8 to 22.7)

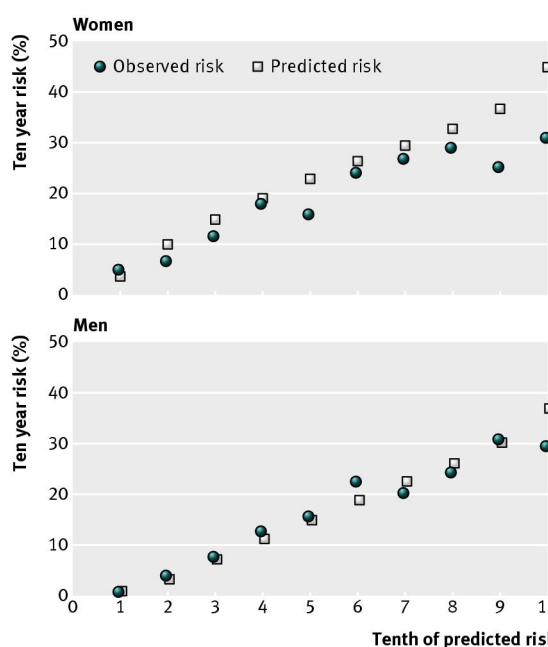
\*For QStroke v CHADS<sub>2</sub>, high risk on QStroke is a predicted 10 year risk  $\geq 15.1\%$ , and high risk on CHADS<sub>2</sub> is score  $\geq 1$ .

†For QStroke v CHA<sub>2</sub>DS<sub>2</sub>VASc, high risk on QStroke is a predicted 10 year risk  $\geq 5.1\%$ , and high risk on CHA<sub>2</sub>DS<sub>2</sub>VASc is score  $\geq 1$ .

## Figures



**Fig 1** Mean predicted risks and observed risk of stroke or transient ischaemic attack at 10 years by tenth of predicted risk applying the QStroke (top) and Framingham (bottom) stroke risk prediction scores to the validation cohort to all patients free of stroke or transient ischaemic attack at baseline



**Fig 2** Mean predicted risks and observed risk of stroke or transient ischaemic attack at 10 years by tenth of predicted risk applying the QStroke risk prediction scores to the subset of patients with atrial fibrillation.