SVN Stroke & Vascular Neurology

Development, validation and comparison of multivariable risk scores for prediction of total stroke and stroke types in Chinese adults: a prospective study of 0.5 million adults

Matthew Chun ,[1,2] Robert Clarke ,[1] Tingting Zhu,[2] David Clifton,[2,3] Derrick A Bennett,[1] Yiping Chen,[1,4] Yu Guo,[5] Pei Pei,[6] Jun Lv,[7,8] Canqing Yu,[7,8] Ling Yang,[1] Liming Li,[7,8] Zhengming Chen,[4] Benjamin J Cairns,[1] On behalf of the China Kadoorie Biobank Collaborative Group

To cite: Chun M, Clarke R, Zhu T, *et al*. Development, validation and comparison of multivariable risk scores for prediction of total stroke and stroke types in Chinese adults: a prospective study of 0.5 million adults. *Stroke & Vascular Neurology* 2022;**7**: e001251. doi:10.1136/svn-2021-001251

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/svn-2021-001251).

RC, ZC and BJC are joint senior authors.

Received 29 July 2021
Accepted 11 January 2022
Published Online First 15 March 2022

For numbered affiliations see end of article.

Correspondence to
Professor Robert Clarke; robert.clarke@ndph.ox.ac.uk

ABSTRACT

Background and purpose Low-income and middle-income countries have the greatest stroke burden, yet remain understudied. This study compared the utility of Framingham versus novel risk scores for prediction of total stroke and stroke types in Chinese adults.

Methods China Kadoorie Biobank (CKB) is a prospective study of 512 726 adults, aged 30–79 years, recruited from 10 areas in China in 2004–2008. By 1 January 2018, 43 234 incident first stroke cases (36 310 ischaemic stroke (IS); 8865 haemorrhagic stroke (HS)) were recorded in 503 842 participants with no history of stroke at baseline. We compared the predictive utility of the Framingham Stroke Risk Profile (FSRP) with novel CKB stroke risk scores and included recalibration, refitting, stratifying by study area and addition of other risk factors. Discrimination was assessed using area under the receiver operating characteristic curve (AUC) and calibration was assessed using Greenwood-Nam-D'Agostino χ^2 statistics.

Results Incidence of total stroke varied fivefold by area in China. The FSRP had good discrimination for total stroke (AUC (95% CI); men: 0.78 (0.77 to 0.79), women: 0.77 (95% CI 0.76 to 0.78)), but poor calibration (χ^2; men: 1,825, women: 3,053), substantially underestimating absolute risks. Recalibration reduced χ^2 by >80%, but did not improve discrimination. Refitting the FSRP did not materially improve discrimination, but further improved calibration. Stratification by area improved discrimination (AUC; men: 0.82 (0.82 to 0.83); women: 0.82 (0.82 to 0.83)), but not calibration. Adding other risk factors yielded modest, but statistically significant, improvements in the AUCs. The findings for IS and HS were similar to those for total stroke.

Conclusions The FSRP reliably differentiated Chinese adults with incident stroke, but substantially underestimated the absolute risks of stroke. Novel local risk prediction equations that took account of differences in stroke incidence within China enhanced risk prediction of total stroke and major stroke pathological types.

INTRODUCTION

Stroke is a leading cause of death and disability worldwide, and about three-quarters of all stroke cases now occur in low-income and middle-income countries (LMICs), including China.[1] Stroke accounted for 34 million prevalent cases and 2 million deaths in China in 2017.[2] Cost-effective primary prevention of stroke requires both population-based lifestyle strategies (eg, salt reduction) and blood pressure-lowering and lipid-lowering medication in high-risk individuals.[3] Risk prediction equations are required to identify those who would derive maximum benefit from such preventive treatments.[4 5]

The Framingham Stroke Risk Profile (FSRP), derived from a multigeneration prospective cohort study in Framingham, Massachusetts, USA, is a widely used risk score for prediction of stroke.[5–8] It provides sex-specific predictions of the absolute risks of total stroke within a specified interval (typically in the next 10 years), based on age, current smoking, history of coronary heart disease (CHD), atrial fibrillation (AF), diabetes, systolic blood pressure and use of antihypertensive treatment.[6–8] Recently updated in 2017, the FSRP has been validated in many high-income countries to predict risk of total stroke,[8 9] but its clinical utility in LMICs, such as China, is uncertain.

The incidence rates of total stroke are higher in China than in Western populations, as are the proportions with haemorrhagic stroke (HS).[10] Within China there are well-documented large, although unexplained differences in the incidence of stroke between geographical areas.[11] Previous studies that estimated absolute risk of total stroke in Chinese

populations were constrained by insufficient numbers of stroke cases, involvement of single rather than multiple areas, lack of reliable information on stroke types (eg, ischaemic stroke (IS) vs HS) and lack of contemporary evidence.[12–14] Consequently, there is a need for more reliable prediction of absolute risks of total stroke and stroke types in Chinese individuals to guide targeted use of evidence-based cost-effective treatments including lipid-lowering and antiplatelet therapy.[15]

Using data from a large prospective study of 0.5M adults recruited into the China Kadoorie Biobank (CKB) in 2004–2008, we compared the performance of the established FSRP with newly developed and internally validated local risk equations to predict the absolute risks of total stroke and stroke types in Chinese adults. The aims of the present report were to develop and validate multivariable risk scores for prediction of total stroke, IS and HS in men and women living in China, and to compare the predictive value of (1) the 2017 FSRP; (2) a recalibrated FSRP; (3) a local recalibrated and refitted FSRP; (4) a recalibrated and refitted FSRP after stratifying by geographical area and (5) area-stratified, recalibrated and refitted models with additional risk factors. A risk calculator for total stroke and stroke pathological types is provided to enable other investigators to validate these stroke risk scores in independent populations.

METHODS

Study population

The data included in the present analyses are available from the corresponding author on reasonable request. Details of the design and methods used in the CKB have been previously reported.[16 17] Briefly, the CKB is a prospective cohort study of 512 726 participants, aged 30–79 years, enrolled from 10 geographically diverse areas (5 urban, 5 rural) of China in 2004–2008. An interviewer-administered electronic questionnaire was used to collect data on sociodemographic factors, lifestyle factors (eg, smoking, alcohol, diet), medical history and current medication and physical activity. Physical measurements included height, weight, hip and waist circumference, bioimpedance, blood pressure and heart rate. All participants provided a blood sample, and random plasma glucose levels were estimated to screen for diabetes. All participants provided written informed consent.

Follow-up for stroke outcomes

The vital status of participants was monitored through death registries supplemented by annual checks with local residential records and active confirmation by contacting local street committees or village administrators.[17] All hospitalised cases of stroke were identified by electronic linkage to established registries of major diseases and health insurance records (covering >97% of participants), supplemented by annual home visits for uninsured participants. All fatal and non-fatal stroke cases were coded by trained medical staff using the International Classification of Diseases 10th revision. The major pathological types of stroke were IS (I63), HS (I60 and I61) and unspecified stroke (I64) (online supplemental eMethods II).[18]

Statistical analyses

The present analyses were restricted to individuals with no prior history of stroke or transient ischaemic attack (205 293 men, 298 549 women) at the date of recruitment. The participants were followed up to detect stroke and death until 1 January 2018, and all incident cases of first stroke (19 587 strokes in men; 23 647 strokes in women) that were recorded for up to 9 years after the baseline survey were included.

For consistency with the sex-specific FSRP and current clinical practice, the present analyses were performed separately in men and women, using time-in-study as the time scale of interest. First, CKB individuals were randomly divided into a training set (85%) and test set (15%). The FSRP was then applied in the test set to predict the risk of total stroke for each individual within 9 years of the baseline survey. Since AF was not recorded in CKB, the FSRP predictions were calculated assuming AF was absent at baseline. No major violations of the proportional hazards assumption for the traditional FSRP covariates were identified (online supplemental eMethods III).

A recalibrated model ('+Recalibration') was subsequently developed, using the Breslow estimator to derive a baseline survival function that adjusted for the mean values of risk factors in CKB,[19 20] while retaining the 2017 FSRP HRs.[8] A recalibrated and refitted model ('+Refitting') was then constructed using Cox regression to derive new HRs for the FSRP risk factors in CKB. For recalibration and refitting, model parameters were derived from the training set, and all models were evaluated using the test set.

To adjust for differences in baseline hazards across the 10 CKB areas, we next developed a model ('+Area stratification') with separate area-specific baselines estimated at the sex-specific mean risk factor values for the overall CKB. In this model, area-stratified Cox regression was used to estimate new HRs for the FSRP risk factors. The model was constructed from the training set and evaluated in the test set.

After estimating separate area-specific baselines, we finally developed an expanded model ('+Additional risk factors') using 133 additional risk indicators recorded at baseline in CKB (online supplemental eWorkbook I), including sociodemographic factors, diet, alcohol consumption, personal and family medical history, physical activity, and physical measurements.[17] The 133 additional risk factors were selected based on their suspected relationship with stroke, while excluding laboratory-based tests, genetic information, and brain imaging that are not widely available in lower-resource clinical settings in China. A subset of these risk factors was then selected automatically using 10-fold cross-validated, least absolute shrinkage and selection operator (LASSO) regularisation (a technique that penalises the inclusion of additional

risk factors to prevent overfitting) within the training set,[21 22] and the selected risk factors were used to fit an area-stratified Cox model using the complete training set. Evaluation of the fitted model was performed using the test set.

Since the associations of individual risk factors with stroke pathological types differ,[15 18 23–25] we also hypothesised that developing separate risk equations for IS and HS could further improve predictive performance compared with a single model for total stroke. Hence, we repeated the analyses separately for IS and for HS pathological types.

To compare predictive performance across models, each model was assessed for discrimination and calibration of 9-year stroke risk predictions using the test set. Risk discrimination refers to the ability to correctly discriminate between individuals with and without stroke, and was evaluated using the area under the receiver operating characteristic curve (AUC). The AUCs for each model were compared with the FSRP using Delong's test.[26] Calibration refers to the similarity between observed and predicted absolute risks and was evaluated using calibration plots. The Greenwood-Nam-D'Agostino χ^2 test statistic was used to compare the observed incidence (calculated as 1 − Kaplan-Meier survival probability) and predicted risks by deciles of predicted risk (with lower χ^2 values indicating better model calibration).[27 28] The 95% CIs were constructed for AUCs using 1000 bootstrapped samples from the test set. For models with area-specific baselines, AUCs were evaluated for both the overall study population and separately within each CKB area.

Sensitivity analyses included restricting the age range to those ≥55 years (for fair comparison with the Framingham study), adding risk factors to the FSRP prior to stratification by study area (to assess the reordering of incremental modelling improvements), and implementing cumulative incidence functions and Fine-Gray models (to account for the competing risk of death from causes other than stroke).[29]

LASSO variable selection and Fine-Gray analyses were performed in R V.3.6.1 using the glmnet package V.3.0–2 and riskRegression package version 8 December 2020, respectively.[21 30] All other statistical analyses were performed using Python V.3.7.0. Cox proportional hazards models were implemented using the lifelines package version 0.21.1.[31] AUC analyses were performed using the scikit-learn toolkit V.0.19.2.[32] Additional details of the methods used for the statistical analyses are provided in online supplemental eMethods IV.

RESULTS

Among the 503 842 CKB study participants in the present analyses, the mean (SD) age was 51.9 (10.6) years and 59% were women. During 9 years of follow-up, a total of 43 234 individuals had a first incident stroke irrespective of type (total stroke); 36 310 had a first IS and 8865 had a first HS (table 1). The incidence of first total stroke was higher in men than in women (9.5% vs 7.9%) and varied over fivefold across the 10 study areas. Compared with those who had no stroke, individuals who had a first stroke were older and more likely to have prior history of

Table 1 Distribution of established risk factors for total stroke and stroke pathological types in men and women in CKB

	Men				Women			
Risk factors included in FSRP*	No stroke (n=185 706)	Total stroke (n=19 587)	IS (n=16 113)	HS (n=4587)	No stroke (n=274 902)	Total stroke (n=23 647)	IS (n=20 197)	HS (n=4278)
Age, mean, year	51.8	60.7	60.8	60.8	50.6	59.6	59.7	59.4
Current smoking, %	68.5	59.9	58.7	63.8	3.1	4.8	4.7	5.4
Coronary heart disease, %	2.1	6.4	6.8	4.8	2.5	9.4	10.0	5.6
Age 65 years+, %	13.8	39.9	39.9	41.4	10.7	34.0	34.1	34.6
Diabetes at age <65 years, %	3.7	6.5	7.0	4.7	4.0	8.1	8.6	6.3
Diabetes at age 65+ years %	1.1	4.8	5.3	3.8	1.3	6.0	6.3	4.7
BP-lowering treatment, %	8.6	22.3	22.6	22.9	10.1	26.3	26.3	29.4
SBP-untreated, mean, mm Hg	130	142	141	148	126	138	137	148
SBP-treated, mean, mm Hg	148	153	152	158	150	155	154	162

'No stroke' column includes individuals lost to follow-up before 9 years, and were stroke-free until being censored.
*Atrial fibrillation was not recorded in CKB.
CKB, China Kadoorie Biobank; FSRP, Framingham Stroke Risk Profile; HS, haemorrhagic stroke; IS, ischaemic stroke; SBP, systolic blood pressure.

CHD, diabetes or hypertension. Individuals who had HS were more likely to be current smokers and have higher mean levels of systolic blood pressure than those who had IS. Overall, men and women had similar rates of prior history of CHD (2.5% vs 3.0%), diabetes (5.3% vs 6.0%), and use of blood pressure-lowering medication (9.9% vs 11.4%), but current smoking was much more common in men than in women (67.7% vs 3.2%) (table 1).

Assessment and update of FSRP for prediction of total stroke

The 2017 FSRP yielded moderate discrimination for total stroke in CKB (AUC (95% CI): 0.78 (0.77 to 0.79) in men, 0.77 (0.76 to 0.78) in women) (table 2). However, calibration was very poor, and the 2017 FSRP substantially underestimated the absolute risk of total stroke (χ^2: 1825 in men, 3053 in women) (table 2; figure 1).

Table 2 Comparison of performance of different models for prediction of total stroke and stroke pathological types in men and women in China Kadoorie Biobank

	Men			Women		
	Discrimination	ΔAUC	Calibration	Discrimination	ΔAUC	Calibration
	AUC (95% CI)	P value	χ^2	AUC (95% CI)	P value	χ^2
Total stroke						
2017 FSRP	0.78 (0.77 to 0.79)	–	1825	0.77 (0.76 to 0.78)	–	3053
+Recalibration	0.78 (0.77 to 0.79)	–	156	0.77 (0.76 to 0.78)	–	506
+Refitting	0.79 (0.79 to 0.80)	+0.01 (<0.001)	51	0.78 (0.77 to 0.78)	+0.01 (<0.001)	148
+Area stratification	0.82 (0.82 to 0.83)	+0.04 (<0.001)	124	0.82 (0.82 to 0.83)	+0.05 (<0.001)	178
+Additional risk factors	0.83 (0.82 to 0.84)	+0.05 (<0.001)	101	0.83 (0.82 to 0.84)	+0.06 (<0.001)	177
Ischaemic stroke						
2017 FSRP	0.77 (0.76 to 0.78)	–	1200	0.76 (0.76 to 0.77)	–	2406
+Recalibration	0.77 (0.76 to 0.78)	–	118	0.76 (0.76 to 0.77)	–	479
+Refitting	0.78 (0.78 to 0.79)	+0.01 (<0.001)	21	0.77 (0.76 to 0.78)	+0.01 (<0.001)	74
+Area stratification	0.82 (0.81 to 0.83)	+0.05 (<0.001)	70	0.82 (0.82 to 0.83)	+0.06 (<0.001)	124
+Additional risk factors	0.83 (0.82 to 0.84)	+0.06 (<0.001)	55	0.83 (0.82 to 0.84)	+0.07 (<0.001)	90
Haemorrhagic stroke						
2017 FSRP	0.79 (0.78 to 0.81)	–	136	0.78 (0.76 to 0.80)	–	70
+Recalibration	0.79 (0.78 to 0.81)	–	58	0.78 (0.76 to 0.80)	–	65
+Refitting	0.80 (0.78 to 0.81)	+0.01 (0.007)	23	0.80 (0.78 to 0.82)	+0.02 (<0.001)	33
+Area stratification	0.81 (0.80 to 0.83)	+0.02 (<0.001)	22	0.81 (0.80 to 0.83)	+0.03 (<0.001)	11
+Additional risk factors	0.82 (0.81 to 0.84)	+0.03 (<0.001)	14	0.82 (0.80 to 0.84)	+0.04 (<0.001)	9

ΔAUC values were calculated as changes from the 2017 FSRP. Modifications to models were applied cumulatively. 2017 FSRP: 2017 FSRP. +Recalibration: Baseline hazard functions re-estimated in China Kadoorie Biobank.+Refitting: Model coefficients from FSRP re-estimated in China Kadoorie Biobank.+Area stratification: Stratification by study area and area-specific estimation of baseline hazard functions.+Additional risk factors: Further sociodemographic, health and lifestyle risk factor indicators selected using LASSO regularisation. AUC, area under the curve; FSRP, Framingham Stroke Risk Profile; LASSO, least absolute shrinkage and selection operator.

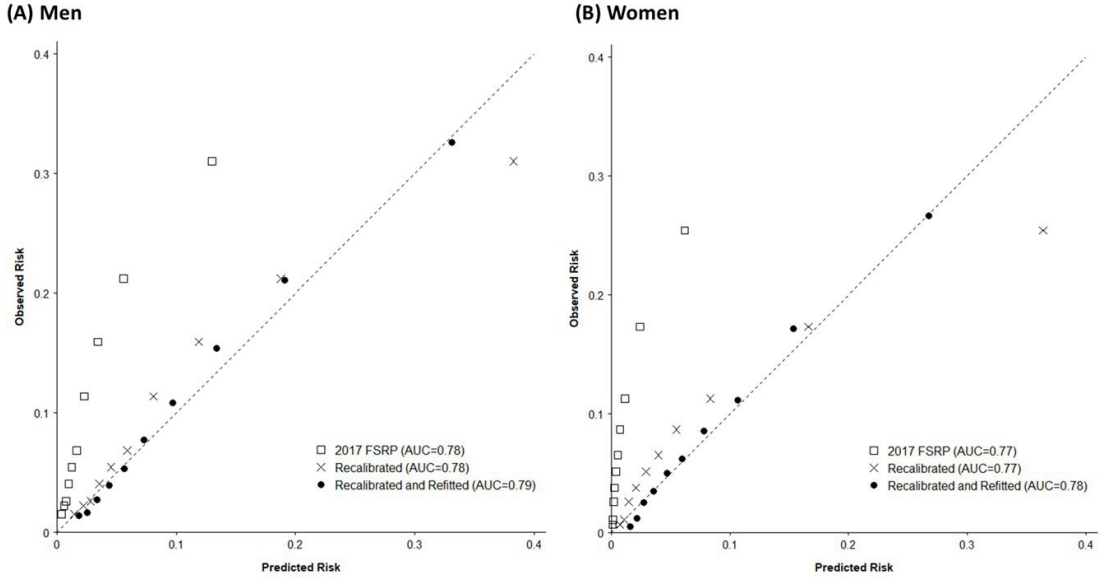


Figure 1 Calibration of the 2017 FSRP, and recalibrated and refitted models from China Kadoorie Biobank, for total stroke in men and women. The dashed line in each subplot represents the line of equality between observed risk and predicted risk. Models with better calibration have points lying closer to the line of equality. Observed 9-year incidence calculated as 1 – Kaplan-Meier estimate. AUC, area under the curve; FSRP, Framingham Stroke Risk Profile.

Recalibration did not alter the AUCs, but substantially corrected the calibration of the model (χ^2: 156 in men, 506 in women). Refitting the HRs for the calibrated equations yielded little material improvement in discrimination (AUC: 0.79 (95% CI 0.78 to 0.80) in men, 0.78 (95% CI 0.77 to 0.78) in women), but further improved calibration (χ^2: 51 in men, 148 in women). Refitted HRs and additional details of these models are provided in online supplemental eWorkbook I.

Prediction of total stroke after adjusting for areas in China

Stroke incidence rates varied markedly by geographical region within China and online supplemental eFigure 1 demonstrates the baseline survival curves for total stroke, IS and HS for each of the 10 study regions in CKB. Modelling separate area-specific baseline survival functions for total stroke yielded modest, but statistically significant improvement (p<0.001) in risk discrimination among all study participants (AUC: 0.82 (95% CI 0.82 to 0.83) in men; 0.82 (95% CI 0.82 to 0.83) in women), while maintaining good calibration (χ^2: 124 in men; 178 in women) (table 2). The discrimination performance within each of the 10 areas is reported in online supplemental eTable I. HRs for individual risk factors obtained from these models differed from the 2017 FSRP (figure 2) and demonstrated substantially greater consistency between men and women and had much greater precision (as reflected by the narrower CIs since the CKB population was 100-fold larger than the Framingham cohort). A sensitivity analysis including age at which ever-regular smokers started smoking had larger HRs associated with ever-regular smoking in men (1.37) and women (1.17), but showed no material improvement in risk prediction for stroke (online supplemental eFigure II).

Expanded risk equations with additional risk factors

In addition to controlling for area-specific differences, the addition of other risk indicators recorded in CKB was assessed for risk prediction of total stroke. The expanded models for total stroke, determined using LASSO regularisation for variable selection, included 66 risk factor indicators for men and 70 in women, including measures of diet, personal and family medical history and socioeconomic status (online supplemental eWorkbook I). These models did not yield any further material improvements in either risk discrimination (AUC: 0.83 (95% CI 0.82 to 0.84) in men, 0.83 (95% CI 0.82 to 0.84) in women) or calibration (χ^2: 101 in men; 177 in women) (table 2). Discrimination performance within each area is reported in the online supplement (online supplemental eTable II).

Risk equations for different stroke pathological types

Analysis of separate risk equations for IS and HS demonstrated comparable results from recalibration, refitting, accounting for geographical area, and addition of other risk factors. The best-performing IS model yielded AUCs (95% CI) of 0.83 (0.82 to 0.84) in men and 0.83 (0.82 to 0.84) in women with χ^2 values of 55 and 90, respectively. The best-performing HS model yielded AUCs of 0.82 (95% CI 0.81 to 0.84) in men and 0.82 (95% CI 0.80 to 0.84) in women with χ^2 values of 14 and 9, respectively (table 2).

The individual risk equations for IS and HS demonstrated substantial differences between the two stroke pathological types. Overall, the absolute risk of IS was 4–5 fold greater than HS, and the ratio of IS to HS risks differed substantially between areas. Modelling area-specific baseline survival curves (ie, predicted survival

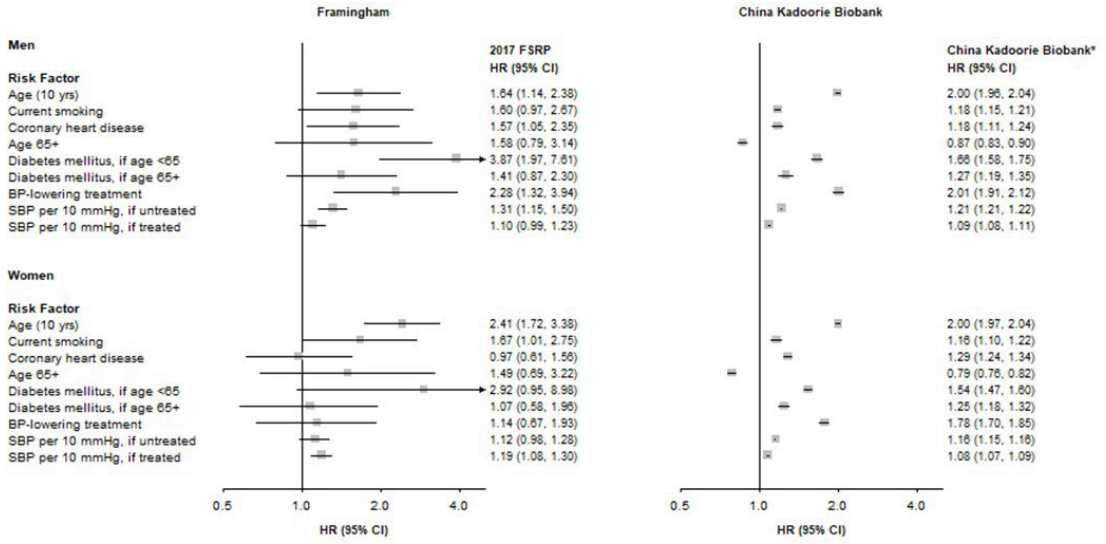


Figure 2 Multivariable HR and 95% CI for total stroke in men and women, for Framingham and for China Kadoorie Biobank. *The 2017 Framingham Stroke Risk Profile (FSRP) coefficients were refitted to China Kadoorie Biobank in a model including stratification by geographical area. BP, blood pressure; SBP, systolic BP.

rates for an individual with mean risk factor values) for IS and HS demonstrated striking differences in stroke risk between areas, consistent with geographical differences in observed stroke incidence during the 9-year follow-up period (figure 3, online supplemental eFigure III). For example, residents in Harbin had threefold higher 9-year

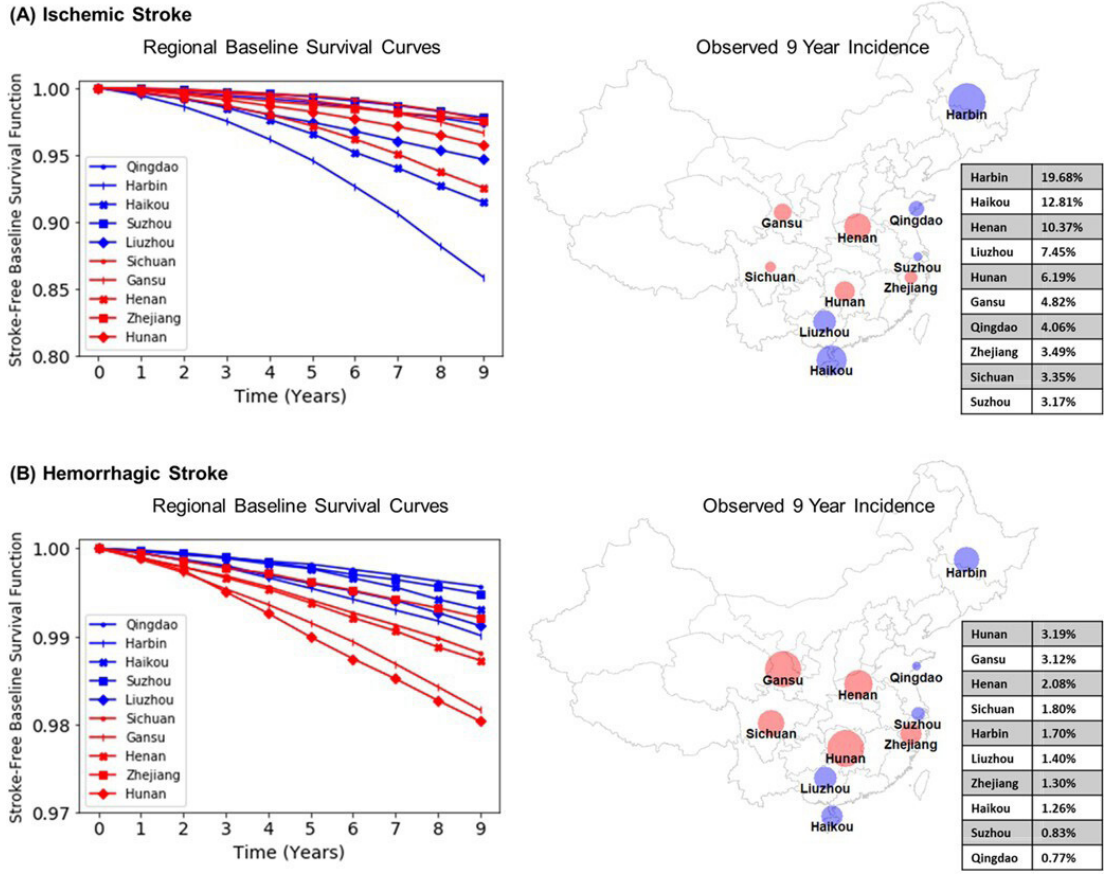


Figure 3 Area-specific baseline survival curves and 9-year incidence for ischaemic stroke and haemorrhagic stroke in China Kadoorie Biobank. Study area-specific baseline survival curves are averaged across sex. Urban study areas are shown in blue while rural areas are shown in red. Note the different scales of the y-axes in subplots (A, B). Dot sizes on maps correspond to observed 9-year incidence (1 – Kaplan-Meier estimate).

incidence of IS compared with those in Hunan (19.68% vs 6.19%), but had half the incidence of HS (1.70% vs 3.19%). Furthermore, by training separate models for IS and HS, different HRs were determined for the same risk factors, including CHD (HR for IS/HS; men: 1.18/1.01, women: 1.32/0.97), diabetes ((age<65 years) men: 1.77/1.44, women: 1.60/1.25; (age 65+years) men: 1.34/1.11, women: 1.30/1.02) and blood pressure-related risk factors (online supplemental eWorkbook I).

Sensitivity analyses

Restriction to a subset of participants aged ≥55 years yielded comparable results from recalibration, refitting, accounting for area and including additional risk factors (online supplemental eTable III). However, due to increased homogeneity among included individuals, the AUCs were lower after excluding the younger adults (age 30–54 years). The age restriction also yielded more extreme differences in the AUCs for the best-performing models for IS and HS compared with the AUCs for the best-performing models for total stroke. Including additional risk factors prior to area-stratification demonstrated that extra risk factors improved AUCs for total stroke, IS and HS, with area-stratification contributing to further improved discrimination for total stroke and IS only (online supplemental eTable IV). After adjusting for competing causes of death, the incidence rates for total stroke and stroke pathological types were similar to the Kaplan-Meier derived estimates (figure 3, online supplemental eFigure IV), and likewise, the predicted risks from the Fine-Gray and Cox models were also similar (online supplemental eFigure V).

Opportunities for validation in independent populations

A risk calculator is provided in online supplemetnal eWorkbook I to enable validation of the CKB risk scores for total stroke and stroke pathological types in independent local populations. Details of the methods on how to use the risk calculators are provided in online supplemental eMethods V.

DISCUSSION

This study, involving a 100-fold larger population than the original Framingham Study, demonstrated that the 2017 FSRP was effective at distinguishing between individuals with and without stroke (good discrimination), but greatly underestimated the absolute risks of total stroke in Chinese adults (poor calibration) due to higher incidence rates of both IS and HS in China compared with Western populations. Absolute risk prediction of total stroke was substantially improved by recalibrating the baseline survival function, with modest additional benefit from refitting HRs (online supplemental eFigure VI). Adjusting for 10 areas in China yielded modest, but statistically significant, improvements in risk discrimination, but there were no further material improvements achieved by adding 38–60 additional risk indicators available in CKB. There was also good performance of separate models for IS and HS, and evidence that the relative importance of predictors differed between these pathological types.

A few population-based prospective studies had previously assessed the utility of FSRP in Chinese adults.[12–14] Overall, they found modest risk discrimination of FSRP for total stroke, but poor prediction of absolute risks of stroke, consistent with the findings of this study. For example, application of FSRP in the China-PAR study, involving 106 281 adults recruited from 4 cohorts in China with a few thousand recorded stroke events, yielded AUCs of 0.65–0.73, but greatly underestimated absolute risks of total stroke.[12] These, and other studies, have highlighted the need for recalibration of Framingham-based equations for prediction of cardiovascular disease in LMICs like China.[33 34] While this study yielded similar findings, it provides several advantages including contemporary risks with much greater precision and reliability due to the very large numbers of well-characterised stroke cases (20-fold greater than the China-PAR study); evaluation of differences by 10 widely distributed geographical areas within China; and separate risk prediction of total stroke, IS and HS.

First, the novel models developed in this study successfully controlled for area-specific differences that were unexplained by analysis of the FSRP risk factors alone. While previous studies such as the China-PAR study have focused on developing a single risk prediction model for the whole country,[12] the results of this study highlight the importance of tailoring risk predictions for specific areas of China, which have substantial differences in incidence of total stroke. The present report provides novel local models for risk prediction of total stroke in 10 diverse areas of China, which have greater predictive utility than a single nationwide model for clinicians in the individual regions.

Second, the separate analysis by study area and stroke pathological types affords insight into the substantial differences in incidence of IS and HS between different areas within China. Some of this geographical variation may be explained by differences in blood pressure.[35] However, much of this variation remains unexplained and may possibly reflect differences in detection (eg, from greater use of brain imaging in certain areas). Inclusion of additional risk factors (eg, sociodemographic factors, alcohol) captured most of the geographical variation in HS risk, but only a fraction for IS risk. Consequently, this study suggests that in studies where explicitly controlling for geographical areas is not feasible, the inclusion of additional risk factors in addition to those included in FSRP could capture some of the regional differences.

Third, this study has significant implications for prevention of different stroke pathological types. Current guidelines in both high-income countries and LMICs advocate the use of blood pressure-lowering medication, lipid-lowering medication and antiplatelet treatment for cardiovascular disease prevention.[36–38] However, individual subtypes of stroke are heterogeneous in their aetiology, and likewise, risk factors have heterogeneous

effects on individual stroke types.[15 18 23–25] This study adds to the available evidence by highlighting differences in the HRs for risk factors such as CHD, diabetes and blood pressure-related variables for IS and for HS, and suggests that evaluating an individual's risk for separate stroke types (as opposed to total stroke only) may also be informative for primary prevention.

This study also had some limitations. First, AF was not recorded in CKB, so could not be included in the models. However, other population-based studies of comparable age groups in China indicated that the prevalence of AF was substantially lower in China than in Framingham (0.4%–1.7% vs 5.0%),[39–41] and in 2012, the AF-related stroke prevalence in China was estimated to be 0.13 per 1000 people.[40] Consequently, although AF is a strong predictor of stroke,[39] it is likely to affect risk prediction for only a small number of individuals in CKB from 2004 to 2008. As prevalence of AF increases in China,[40 41] it may be increasingly important to incorporate AF into future local stroke risk equations. Another limitation of CKB was that recorded stroke events were limited to hospitalised strokes (92% having brain imaging to support diagnosis) and death from stroke.[18] In addition, while the risk equations presented in this study are useful for risk prediction, the HRs for individual risk factors cannot be interpreted causally.

Moreover, the Cox models presented do not account for competing risks of death due to other causes, which may affect risk estimates, particularly in older individuals. However, with low rates of censoring in CKB (5.4%, with 4.8% of censoring due to death), the effect of this limitation is small, as indicated by the comparable 9-year stroke incidence rates of stroke after adjusting for competing risk of death and the similar predicted risks between the Cox and Fine-Gray models.

Finally, the risk equations outlined in the present report were not designed for immediate implementation in clinical practice, which would require additional validation in independent populations in China and potentially other LMICs. To our knowledge, there are currently no contemporary regional cohorts of middle-aged and older adults with sufficiently large sample size in China to perform an external validation. As such datasets become available (eg, via the China Precision Medicine Initiative and establishment of regional electronic health records), future studies can use the calculator provided (online supplemental eWorkbook I) to validate these equations.

CONCLUSIONS

This study developed novel local risk equations for total stroke, IS and HS and demonstrated modest, but statistically significant, improvements over the widely used 2017 FSRP in Chinese adults. Improvements in stroke risk prediction can be attributed to recalibration of baseline survival, refitting HRs and accounting for geographical differences in stroke incidence in China. The addition of a large number of other risk factors yielded no further material improvements, but may be useful in other studies when area-specific differences are not readily estimated. These techniques can be implemented to improve risk prediction in any Chinese or similar populations with unique and geographically diverse risk profiles for stroke. Moreover, separate risk equations for IS and HS could help to identify individuals at high risk of a particular stroke pathological type and guide treatment decisions for primary prevention. These equations should be validated and refined in independent populations before implementing them for prediction of stroke risk in clinical practice.

Author affiliations
[1]Clinical Trial Service Unit and Epidemiological Studies, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[2]Department of Engineering Science, University of Oxford, Oxford, UK
[3]Department of Biomedical Engineering, Oxford-Suzhou Centre for Advanced Research, Suzhou, China
[4]Medical Research Council Health Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[5]CKB Project Department, Fuwai Hospital Chinese Academy of Medical Sciences, National Center for Cardiovascular Diseases, Beijing, China
[6]CKB Project Department, Chinese Academy of Medical Sciences, Beijing, China
[7]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China
[8]Department of Epidemiology, Peking University Center for Public Health and Epidemic Preparedness and Response, Beijing, China

Acknowledgements The chief acknowledgment is to the participants, the project staff, and the China National Centre for Disease Control and Prevention and its regional offices for access to death and disease registries. The Chinese National Health Insurance scheme provides electronic linkage to all hospital admission data.

Collaborators China Kadoorie Biobank Collaborative Study Group.

Contributors Study concept and design: MC, TZ, DC, BJC and RC. Data collection: RC, DAB, YG, YC, PP, JL, CY, LY, LL and ZC. Data analysis and interpretation: MC, TZ, DC, BJC and RC. Drafting of the manuscript: MC, TZ, DC, BJC and RC. Critical revision of the manuscript: all authors. Final approval: all authors.

Funding The baseline survey was funded by the Kadoorie Charitable Foundation, Hong Kong, China and the funding sources for the long-term continuation of the study included UK Wellcome Trust (202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z), Chinese National Natural Science Foundation (81390540, 81390541, 81390544) and the National Key Research and Development Program of China (2016YFC0900500, 2016YFC0900501, 2016YFC0900504, 2016YFC1303904). Core funding was also provided to the CTSU, University of Oxford, by the British Heart Foundation (CH/1996001/9454), the UK Medical Research Council, and Cancer Research UK. MC was supported by a Rhodes Scholarship. BJC was supported by a Nuffield Department of Population Health Senior Research Fellowship. The University of Oxford Medical Research Council (MRC) Population Health Research Unit is funded through a strategic partnership between the MRC and the University of Oxford (MC_UU_00017/1, MC_UU_12026/2, MC_U137686851). The research was also supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Oxford Tropical Research Ethics Committee (OxTREC) (2005)Reference number (OxTREC): 025-04.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Access details to a stroke risk calculator are provided in a workbook in the online supplemental materials to enable researchers to calculate risk scores for stroke using their own data. Researchers who are interested in obtaining the raw data from the China Kadoorie Biobank study that underlines this paper should contact ckbaccess@ndph.ox.ac.uk. A research proposal will be requested to ensure that any analysis is performed by bona fide researchers and - where data is not currently available to open access researchers - is restricted to the topic covered in this paper.

ORCID iDs
Matthew Chun http://orcid.org/0000-0002-8819-2072
Robert Clarke http://orcid.org/0000-0002-9802-8241

REFERENCES

1 Feigin VL, Krishnamurthi RV, Parmar P, *et al*. Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: the GBD 2013 study. *Neuroepidemiology* 2015;45:161–76.
2 Institute for Health Metrics and Evaluation. *GBD compare*. Seattle, WA: IHME, University of Washington, 2015. https://vizhub.healthdata.org/gbd-compare
3 Liu M, Wu B, Wang W-Z, *et al*. Stroke in China: epidemiology, prevention, and management strategies. *Lancet Neurol* 2007;6:456–64.
4 Richards A, Cheng EM. Stroke risk calculators in the era of electronic health records linked to administrative databases. *Stroke* 2013;44:564–9.
5 Meschia JF, Bushnell C, Boden-Albala B, *et al*. Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American heart Association/American stroke association. *Stroke* 2014;45:3754–832.
6 Wolf PA, D'Agostino RB, Belanger AJ, *et al*. Probability of stroke: a risk profile from the Framingham study. *Stroke* 1991;22:312–8.
7 D'Agostino RB, Wolf PA, Belanger AJ, *et al*. Stroke risk profile: adjustment for antihypertensive medication. The Framingham study. *Stroke* 1994;25:40–3.
8 Dufouil C, Beiser A, McLure LA, *et al*. Revised Framingham stroke risk profile to reflect temporal trends. *Circulation* 2017;135:1145–59.
9 Flueckiger P, Longstreth W, Herrington D, *et al*. Revised Framingham stroke risk score, nontraditional risk markers, and incident stroke in a multiethnic cohort. *Stroke* 2018;49:363–9.
10 Tsai C-F, Thomas B, Sudlow CLM. Epidemiology of stroke and its subtypes in Chinese vs white populations: a systematic review. *Neurology* 2013;81:264–72.
11 Chen JS, Campbell TC, JY L, *et al*. *Lifestyle and mortality in China: a study of the characteristics of 65 Chinese counties*. Oxford University Press, 1990.
12 Xing X, Yang X, Liu F, *et al*. Predicting 10-year and lifetime stroke risk in Chinese population. *Stroke* 2019;50:2371–8.
13 Chien K-L, Su T-C, Hsu H-C, *et al*. Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan. *Stroke* 2010;41:1858–64.
14 Leung JY, Lin SL, Lee RS, *et al*. Framingham risk score for predicting cardiovascular disease in older adults in Hong Kong. *Hong Kong Med J* 2018;24(Suppl 4):S8–11.
15 Sun L, Clarke R, Bennett D, *et al*. Causal associations of blood lipids with risk of ischemic stroke and intracerebral hemorrhage in Chinese adults. *Nat Med* 2019;25:569–74.
16 Chen Z, Lee L, Chen J, *et al*. Cohort profile: the Kadoorie study of chronic disease in China (KSCDC). *Int J Epidemiol* 2005;34:1243–9.
17 Chen Z, Chen J, Collins R, *et al*. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* 2011;40:1652–66.
18 Chen Y, Wright N, Guo Y, *et al*. Mortality and recurrent vascular events after first incident stroke: a 9-year community-based study of 0·5 million Chinese adults. *Lancet Glob Health* 2020;8:e580–90.
19 Breslow NE. Discussion of the paper by DR. Cox. *J R Statist Soc B* 1972;34:215–6.
20 Lin DY. On the Breslow estimator. *Lifetime Data Anal* 2007;13:471–80.
21 Friedman J, Hastie T, Tibshirani R. Glmnet: lasso and elastic-net regularized generalized linear models. R package version 3.0-2, 2019. Available: http://CRAN.R-project.org/package=glmnet [Accessed 8 Apr 2020].
22 Hastie T, Qian J. Glmnet vignette, 2014. Available: http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf [Accessed 8 Apr 2020].
23 Chen Z, Iona A, Parish S, *et al*. Adiposity and risk of ischaemic and haemorrhagic stroke in 0·5 million Chinese men and women: a prospective cohort study. *Lancet Glob Health* 2018;6:e630–40.
24 Holmes MV, Millwood IY, Kartsonaki C, *et al*. Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J Am Coll Cardiol* 2018;71:620–32.
25 Dong W, Pan X-F, Yu C, *et al*. Self-rated health status and risk of incident stroke in 0.5 million Chinese adults: the China Kadoorie Biobank study. *J Stroke* 2018;20:247–57.
26 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
27 D'Agostino RB, Byung-Ho N. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of Statistics* 2004;23:1–25.
28 Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. *Stat Med* 2015;34:1659–80.
29 Fine JP, Gray RJ. A proportional hazards model for the Subdistribution of a competing risk. *J Am Stat Assoc* 1999;94:496–509.
30 Gerds TA, Blanche P, Mortensen R. riskRegression: risk regression models and prediction scores for survival analysis with competing risks. R package version 2020.12.08, 2020. Available: http://CRAN.R-project.org/package=riskRegression [Accessed 12 Apr 2021].
31 Davidson-Pilon C, Kalderstam J, Jacobson N. CamDavidsonPilon/lifelines: v0.21.1 (version v0.21.1). Zenodo, 2019. Available: http://doi.org/10.5281/zenodo.2652543
32 Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in python. *JMLR* 2011;12:2825–30.
33 Damen JA, Pajouheshnia R, Heus P, *et al*. Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis. *BMC Med* 2019;17:109.
34 Pennells L, Kaptoge S, Wood A, *et al*. Equalization of four cardiovascular risk algorithms after systematic recalibration: individual-participant meta-analysis of 86 prospective studies. *Eur Heart J* 2019;40:621–31.
35 Lewington S, Li L, Sherliker P, *et al*. Seasonal variation in blood pressure and its relationship with outdoor temperature in 10 diverse regions of China: the China Kadoorie Biobank. *J Hypertens* 2012;30:1383–91.
36 Arnett DK, Blumenthal RS, Albert MA, *et al*. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American heart association Task force on clinical practice guidelines. *Circulation* 2019;140:e596–646.
37 Joint Committee for Guideline Revision. 2018 Chinese guidelines for prevention and treatment of Hypertension-A report of the revision Committee of Chinese guidelines for prevention and treatment of hypertension. *J Geriatr Cardiol* 2019;16:182–241.
38 Joint Committee for Guideline Revision. 2016 Chinese guidelines for the management of dyslipidemia in adults. *J Geriatr Cardiol* 2018;15:1–29.
39 Zhou Z, Hu D. An epidemiological study on the prevalence of atrial fibrillation in the Chinese population of mainland China. *J Epidemiol* 2008;18:209–16.
40 Guo Y, Tian Y, Wang H, *et al*. Prevalence, incidence, and lifetime risk of atrial fibrillation in China: new insights into the global burden of atrial fibrillation. *Chest* 2015;147:109–19.
41 Wang Z, Chen Z, Wang X, *et al*. The disease burden of atrial fibrillation in China from a national cross-sectional survey. *Am J Cardiol* 2018;122:793–8.

ONLINE SUPPLEMENTAL FILE 1

Risk scores for prediction of total stroke and stroke types in Chinese adults: a prospective study of 0.5 million adults

Supplementary Methods

Supplementary Figures

Supplementary Tables

Supplementary Files

Online supplemental file 2. Stroke Risk Calculator (Including All CKB Risk Factors, Selected Risk Factors for Each Stroke Risk Equation, and Associated Hazard Ratios)

eMethods I: Members of the China Kadoorie Biobank Collaborative Group

International Steering Committee: Junshi Chen, Zhengming Chen (PI), Robert Clarke, Rory Collins, Yu Guo, Liming Li (PI), Chen Wang, Jun Lv, Richard Peto, Robin Walters.

International Co-ordinating Centre, Oxford: Daniel Avery, Derrick Bennett, Ruth Boxall, Ka Hung Chan, Yumei Chang, Yiping Chen, Zhengming Chen, Johnathan Clarke; Robert Clarke, Huaidong Du, Zammy Fairhurst-Hunter, Hannah Fry, Simon Gilbert, Alex Hacker, Mike Hill, Michael Holmes, Pek Kei Im, Andri Iona, Maria Kakkoura, Christiana Kartsonaki, Rene Kerosi, Kuang Lin, Mohsen Mazidi, Iona Millwood, Qunhua Nie, Alfred Pozarickij, Paul Ryder, Saredo Said, Sam Sansome, Dan Schmidt, Paul Sherliker, Rajani Sohoni, Becky Stevens, Iain Turnbull, Robin Walters, Lin Wang, Neil Wright, Ling Yang, Xiaoming Yang, Pang Yao.

National Co-ordinating Centre, Beijing: Yu Guo, Xiao Han, Can Hou, Qingmei Xia, Chao Liu, Jun Lv, Pei Pei, Canqing Yu.

10 Regional Co-ordinating Centres: Guangxi Provincial CDC: Naying Chen, Duo Liu, Zhenzhu Tang. Liuzhou CDC: Ningyu Chen, Qilian Jiang, Jian Lan, Mingqiang Li, Yun Liu, Fanwen Meng, Jinhuai Meng, Rong Pan, Yulu Qin, Ping Wang, Sisi Wang, Liuping Wei, Liyuan Zhou. **Gansu** Provincial CDC: Caixia Dong, Pengfei Ge, Xiaolan Ren. Maiji CDC: Zhongxiao Li, Enke Mao, Tao Wang, Hui Zhang, Xi Zhang. **Hainan** Provincial CDC: Jinyan Chen, Ximin Hu, Xiaohuan Wang. Meilan CDC: Zhendong Guo, Huimei Li, Yilei Li, Min Weng, Shukuan Wu. **Heilongjiang** Provincial CDC: Shichun Yan, Mingyuan Zou, Xue Zhou. Nangang CDC: Ziyan Guo, Quan Kang, Yanjie Li, Bo Yu, Qinai Xu. **Henan** Provincial CDC: Liang Chang, Lei Fan, Shixian Feng, Ding Zhang, Gang Zhou. Huixian CDC: Yulian Gao, Tianyou He, Pan He, Chen Hu, Huarong Sun, Xukui Zhang. **Hunan** Provincial CDC: Biyun Chen, Zhongxi Fu, Yuelong Huang, Huilin Liu, Qiaohua Xu, Li Yin. Liuyang CDC: Huajun Long, Xin Xu, Hao Zhang, Libo Zhang. **Jiangsu** Provincial CDC: Jian Su, Ran Tao, Ming Wu, Jie Yang, Jinyi Zhou, Yonglin Zhou. Suzhou CDC: Yihe Hu, Yujie Hua, Jianrong Jin Fang Liu, Jingchao Liu, Yan Lu, Liangcai Ma, Aiyu Tang, Jun Zhang. **Qingdao** Qingdao CDC: Liang Cheng, Ranran Du, Ruqin Gao, Feifei Li, Shanpeng Li, Yongmei Liu, Feng Ning, Zengchang Pang, Xiaohui Sun, Xiaocao Tian, Shaojie Wang, Yaoming Zhai, Hua Zhang, Licang CDC: Wei Hou, Silu Lv, Junzheng Wang. **Sichuan** Provincial CDC: Xiaofang Chen, Xianping Wu, Ningmei Zhang, Weiwei Zhou. Pengzhou CDC: Xiaofang Chen, Jianguo Li, Jiaqiu Liu, Guojin Luo, Qiang Sun, Xunfu Zhong. **Zhejiang** Provincial CDC: Weiwei Gong, Ruying Hu, Hao Wang,Meng Wan, Min Yu. Tongxiang CDC: Lingli Chen, Qijun Gu, Dongxia Pan, Chunmei Wang, Kaixu Xie, Xiaoyi Zhang.

Event Adjudication Clinicians: Beijing Tiantan Hospital,Capital Medical University Shuya Li, Haiqiang Qin, Yongjun Wang, **Peking University People's Hospital** Qiling Chen, Jihua Wang, **The 1st Affiliated Hospital of Harbin Medical University** Xiaojia Sun, Lei Wang, Xun Wang, Liming Zhang, Shanshan Zhou, **The 2nd Affiliated Hospital of Harbin Medical University** Hongyuan Chen, Li Chen, Haiyan Gou, Weizhi Wang, Yanmei Zhu, Yulan Zhu, **The 2nd Hospital of Hebei Medical University** Ning Zhang, **Huashan Hospital** Xin Cheng, Qiang Dong, Yi Dong, Kun Fang, Yiting Mao, **Jinling Hospital** Yu An, Peiling Chen, Yinghua Chen, Zhihong Liu, Lihua Zhang **The People's Hospital of Liaoning Province** Xiaohong Chen, Naixin Jv, Xiaojiu Li, Liyang Liu, Yun Lu, Xiaona Xing, **Qingdao Fuwai Cardiovascular**

Hospital Shihao You, **Shengjing Hospital of China Medical University** Xiaoli Cheng, Chaojun Gua, Jinping Jiang, Jingyi Liu, Shumei Ma, **Shenyang Military General Hospital** Xuefeng Yang, **The First People's Hospital of Shenyang** Xiaomo Du, Jian Xu, Xuecheng Yang, Xiaodi Zhao, **West China Hospital, Sichuan University** Zilong Hao, Ming Liu, Deren Wang, **The Second Affiliated Hospital of Suzhou University** Xiaoting Li, **Suzhou Kowloon Hospital Shanghai Jiao Tong University School of Medicine** Lili Hui, Zhanling Liao, Feng Liu, **Qingdao Fuwai Cardiovascular Hospital** Chunning Feng, Dejiang Ji, Fengxia Qu, Wenwen Yuan, **The First Affiliated Hospital of Zhengzhou University** Xin Fu, **Zhongshan Hospital,** Jing Ding, Peng Du, Lirong Jin, Yueshi Mao, Xin Wang.

eMethods II. Definitions of Stroke Types

Ischemic stroke (ICD-10 code I63), including lacunar infarction and non-lacunar infarction, was defined as a focal neurological dysfunction lasting for more than 24 hours with or without neuroimaging evidence of a cerebral infarct.

Hemorrhagic stroke was defined to include intracerebral hemorrhage (ICD-10 code I61) and subarachnoid hemorrhage (ICD-10 code I60). Intracerebral hemorrhage was defined as neurological dysfunction caused by hemorrhage into the brain parenchyma or the ventricular system, excluding those induced by injury, with or without neuroimaging evidence of brain hemorrhage. Subarachnoid hemorrhage was defined as neurological dysfunction caused by hemorrhage into the subarachnoid space, excluding those induced by injury, with or without neuroimaging evidence of such hemorrhage.

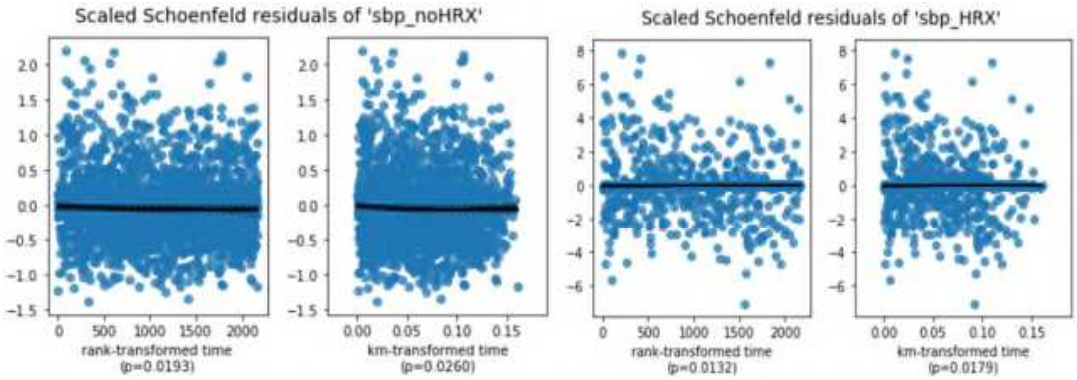
All fatal and non-fatal stroke cases were coded using ICD-10 by trained medical staff, who were blinded to other personal information, with further checking and review conducted centrally by trained medical staff. All hospital-reported cases of first stroke also underwent additional clinical adjudication, involving retrieval and review of original medical records and brain imaging reports by clinical specialists in China using a bespoke web-based system. About 92% of the reported first stroke cases had their diagnosis confirmed by brain imaging (CT or MRI). Radiological reports (but not primary brain images) of reported cases of non-fatal stroke were adjudicated by Chinese neurologists using a bespoke online system.[1]

eMethods III. Testing of Proportional Hazards Assumption

The Proportional Hazards (PH) assumption for the FSRP inputs was checked using the Cox PHFitter.check_assumptions method implemented by the lifelines package[2] version 0.21.1 in Python version 3.7.0. This method performs a statistical test to test for any time-varying coefficients, and provides visual plots of the scaled Schoenfeld residuals presented against four time transformations for any risk factor that violates the PH assumption. In each plot, a fitted lowess is also presented, along with 10 bootstrapped lowess lines. Deviations of the lowess line from a constant value are violations of the PH assumption.

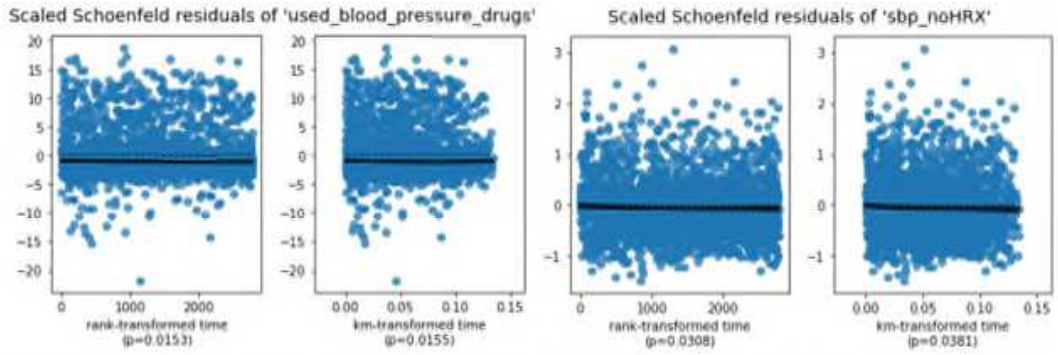
Tests of the PH assumption were performed for each FSRP risk factor and separately by sex for CKB participants in the training set (174,499 men; 253,766 women). In CKB men, anti-hypertensive treatment ($p<0.005$), systolic blood pressure for individuals without hypertension treatment ($p<0.005$), and systolic blood pressure for individuals with hypertension treatment ($p=0.02$) were identified to violate the PH assumption (using a p-value threshold of 0.05). In CKB women, the risk factors identified to violate the PH assumption were age ($p<0.005$), diabetes if under 65 years ($p<0.005$), diabetes if 65+ years ($p=0.01$), anti-hypertensive treatment ($p<0.005$), systolic blood pressure for individuals without hypertension treatment ($p<0.005$), and systolic blood pressure for individuals with hypertension treatment ($p=0.04$).

It is important to note that with a large sample size, such as in CKB, even very small violations of the Proportional Hazards Assumption would test as statistically significant. To observe the impact of this effect, we repeated tests of the PH assumption for a randomly selected 10% of CKB men and women (17,449 men; 25,376 women). In men, only systolic blood pressure for individuals without hypertension treatment ($p=0.02$) and systolic blood pressure for individuals with hypertension treatment ($p=0.01$) were still identified to violate the PH assumption. However, even among these risk factors, the Schoenfeld residual plots (below) showed a lowess line with very minor deviation from a constant value.



Schoenfeld residual plots for systolic blood pressure for individuals without hypertension treatment (left 2 subplots) and systolic blood pressure for individuals with hypertension treatment (right 2 subplots) among CKB men

In women, only anti-hypertensive treatment ($p=0.02$) and systolic blood pressure for individuals without hypertension treatment ($p=0.03$) were still identified to violate the PH assumption. Once again, the Schoenfeld residual plots for these risk factors (below) showed a lowess line with very minor deviation from a constant value. The results of these tests suggest that while the hazards for every FSRP risk factor may not be perfectly proportional, the PH assumption may still be appropriate in this setting.



Schoenfeld residual plots for anti-hypertensive treatment (left 2 subplots) and systolic blood pressure for individuals without hypertension treatment (right 2 subplots) among CKB women

Finally, we note that while the PH assumption is important for causal inference, satisfaction of the PH assumption is not necessary for risk prediction applications, where the objectives are optimal risk discrimination and risk calibration irrespective of how individual predictions are generated.

eMethods IV. Model Development and Statistical Analyses

Training Set and Test Set Split

The CKB data was divided into a training set and test set using a random 85%/15% training/test split, stratified by occurrence of the relevant endpoint (i.e., total stroke, IS, or HS) within 9 years of the baseline survey.

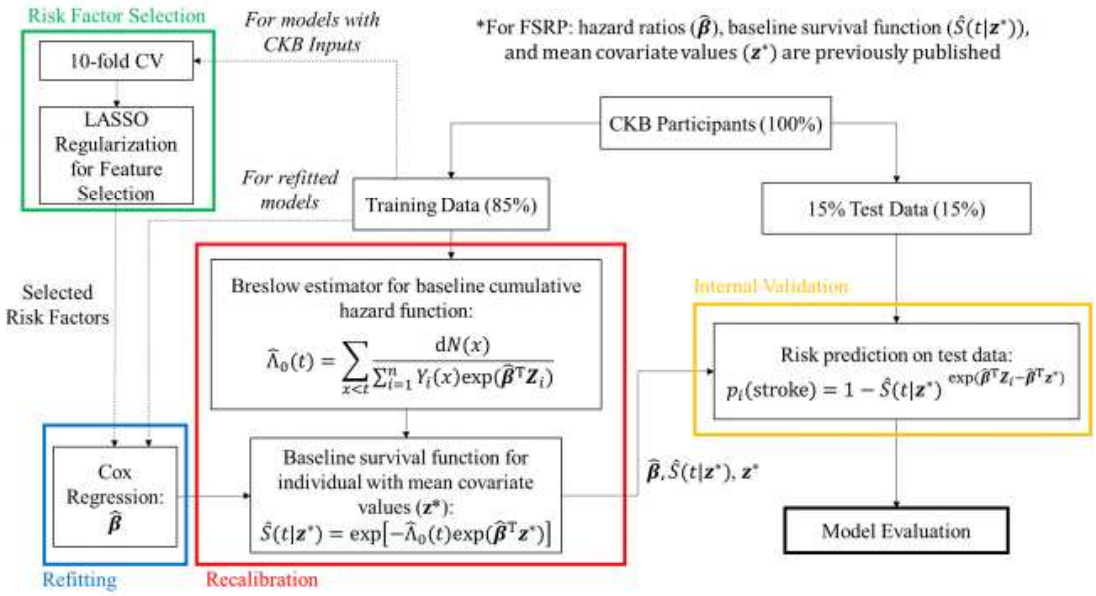
Missing Values

Missing values in both the training set and test set were imputed using the means of the non-missing values in the training set. CKB has very few missing values. Of the 133 risk factors considered in CKB (listed with definitions in eWorkbook 1), only 16 risk factors had missing values. Out of all 503,842 individuals included in the present analyses, 521 individuals had missing values for number of siblings and siblings' medical history (stroke, heart attack, diabetes, and cancer); 817 had missing values for mother's medical history (stroke, heart attack, diabetes, and cancer); 1,124 had missing values for father's medical history (stroke, heart attack, diabetes, and cancer); 2 had missing values for weight and BMI; and 228 had missing values for body fat percentage. In addition to mean imputation, 3 binary risk factors were added to represent whether or not an individual was missing medical history for their mother, father, or siblings, respectively. This resulted in a total of 133 potential CKB inputs.

Model Construction

All of the risk equations evaluated in the present analyses were developed using the Cox proportional hazards regression model. Hazard ratios for risk factors were derived from Cox regression while the baseline hazard function (and corresponding baseline survival function) were derived from the non-parametric Breslow estimator.[3,4]

For model construction, recalibration of the baseline survival function and refitting of hazard ratios were conducted in the training set. A summary of the recalibration and refitting methodology is provided in the following diagram, and further details are provided below.



Aggregate Baseline w/ FSRP Inputs: For the following models, a single "aggregate" baseline survival function was implemented for all individuals in the study population. We describe this as an "aggregate" baseline because the survival function was based on a combination of all individuals regardless of their area.

2017 FSRP ("2017 FSRP"): For these models, the published 2017 FSRP baseline survival function and FSRP hazard ratios were used without performing any recalibration or refitting procedures.

Recalibrated FSRP ("+ Recalibration"): For these models, an aggregate CKB baseline survival function was derived. 2017 FSRP hazard ratios were used without refitting.

Recalibrated and Refitted FSRP ("+ Refitting"): For these models, an aggregate CKB baseline survival function was derived and new hazard ratios were refitted for the FSRP risk factors.

Area-specific Baselines w/ FSRP Inputs ("+ Area stratification"): For these models, separate baseline survival functions were developed for each CKB area (10 in total) and new hazard ratios were refitted for the FSRP risk factors using an area-stratified Cox model.

Area-specific Baselines w/ CKB Inputs ("+ Additional risk factors"): For these models, separate baseline survival functions were developed for each CKB area (10 in total). However, rather than limiting the model to FSRP inputs, 10-fold cross-validated LASSO regularization was used (within the training set) for selecting a subset of risk factors from all CKB variables. LASSO regularization was performed separately for each model, yielding slightly different numbers of selected risk factors. The specifics of the variable selection process have been previously described.[5,6] Hazard ratios for the selected risk factors were then fitted using an area-stratified Cox model on the complete training set.

Risk Equations for IS and HS: For these models, identical procedures were performed as described above, replacing total stroke with IS and HS. However, since the FSRP is only provided as a total stroke model, the 2017 FSRP baseline survival function and hazard ratios for total stroke were also used for the IS and HS risk equations.

Model Evaluation

All models were internally validated using the test set. Risk discrimination was evaluated using the area under the receiver operating characteristic curve (AUC), and calibration was evaluated using the Greenwood-Nam-D'Agostino chi-squared statistic ($\chi 2$). Mean values and 95% confidence intervals were determined using 1000 bootstrapped samples from the test set. For comparison, the training set, test set, and bootstrapped samples were designed to be identical for all models predicting the same endpoint (i.e, all total stroke models, all ischemic stroke models, and all hemorrhagic stroke models).

Model Reporting

In order to avoid over-optimism, all reporting about the performance of the risk prediction equations, including AUC and $\chi 2$ metrics, were based on the performance of these models in the test set only.

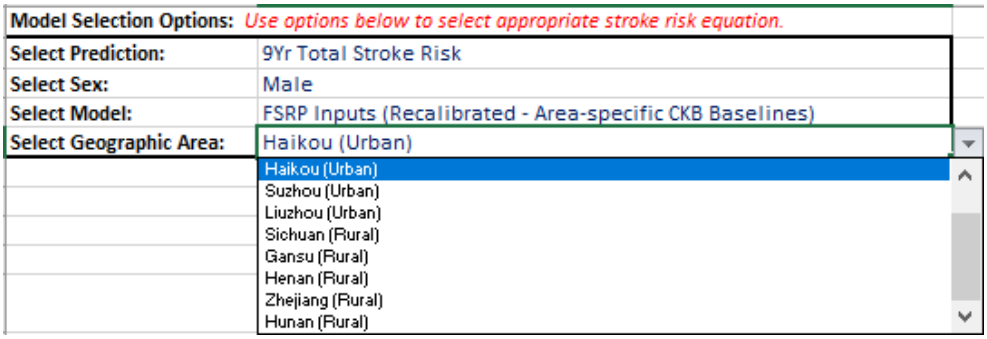
However, in order to capture information from all CKB individuals in our final models, we report hazard ratios in Figure 2 and eFigure II, risk predictions in eFigure III, and full risk equations in eWorkbook I after reconstructing each model using the overall study population.

eMethods V. Procedures for Applying Reported Models to Predict 9-Year Risk of First Total Stroke, First Ischemic Stroke, or First Hemorrhagic Stroke

Caution: The models presented in this report should not be used in clinical practice before being validated and refined in independent populations. These procedures are included for reporting transparency and reproducibility purposes only.

Calculating a Stroke Risk Prediction for an Individual

We include an interactive tool for exploring the risk equations reported in this study. To determine an individual's 9-yr predicted risk of stroke, refer to the attached "Stroke Risk Calculator" (**Online supplemental file 2**). Using the "Model Selection Options" section in columns A and B, use the provided dropdown menus to select the model of interest.

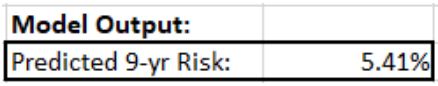


- *Select Prediction*: Allows user to select between "9Yr Total Stroke Risk", "9Yr Ischemic Stroke Risk", and "9Yr Hemorrhagic Stroke Risk"
- *Select Sex*: Allows user to choose between "Male and "Female" options.
- *Select Model*: Allows user to select model-of-interest from the present study
- *Select Geographic Area*: Allows user to select area of individual. If the selected model-of-interest does not require the individual's area, "N/A" will be the only available option.

After completing the "Model Selection Options", refer to columns D and E for the "Individual Risk Factor Values" section. Depending on the selected model, corresponding risk factor prompts will appear in column D. Enter the individual's risk factor values in column E. If any prompts are cut-off, you may need to rewrap the text in the cell. This can be done by highlighting the relevant cell and toggling the "wrap text" button in the "Home" tab of Excel.

Individual Risk Factor Values:	
Risk Factors:	*Insert risk factor values for individual in this column*
Age (yrs)	40
Current Smoking (0 for no; 1 for yes)	0
History of CHD? (0 for no; 1 for yes)	0
Has Diabetes? (0 for no; 1 for yes)	0
Used BP-lowering drugs? (0 for no; 1 for yes)	0
SBP (in mmHg)	130

Once all risk factor values are entered, a calculated predicted 9-yr risk will be presented in the "Model Output" section in columns G and H.



Model Output:	
Predicted 9-yr Risk:	5.41%

Additional Information on Reported Models

The worksheet titled "Model Params - HRs" includes a full listing of all CKB variables, the selected risk factors used in each model (displayed in green), and their corresponding hazard ratios.

The worksheet titled "Model Params - Base Surv" includes the baseline survival function, evaluated at 9 years from time-of-prediction, for each model.

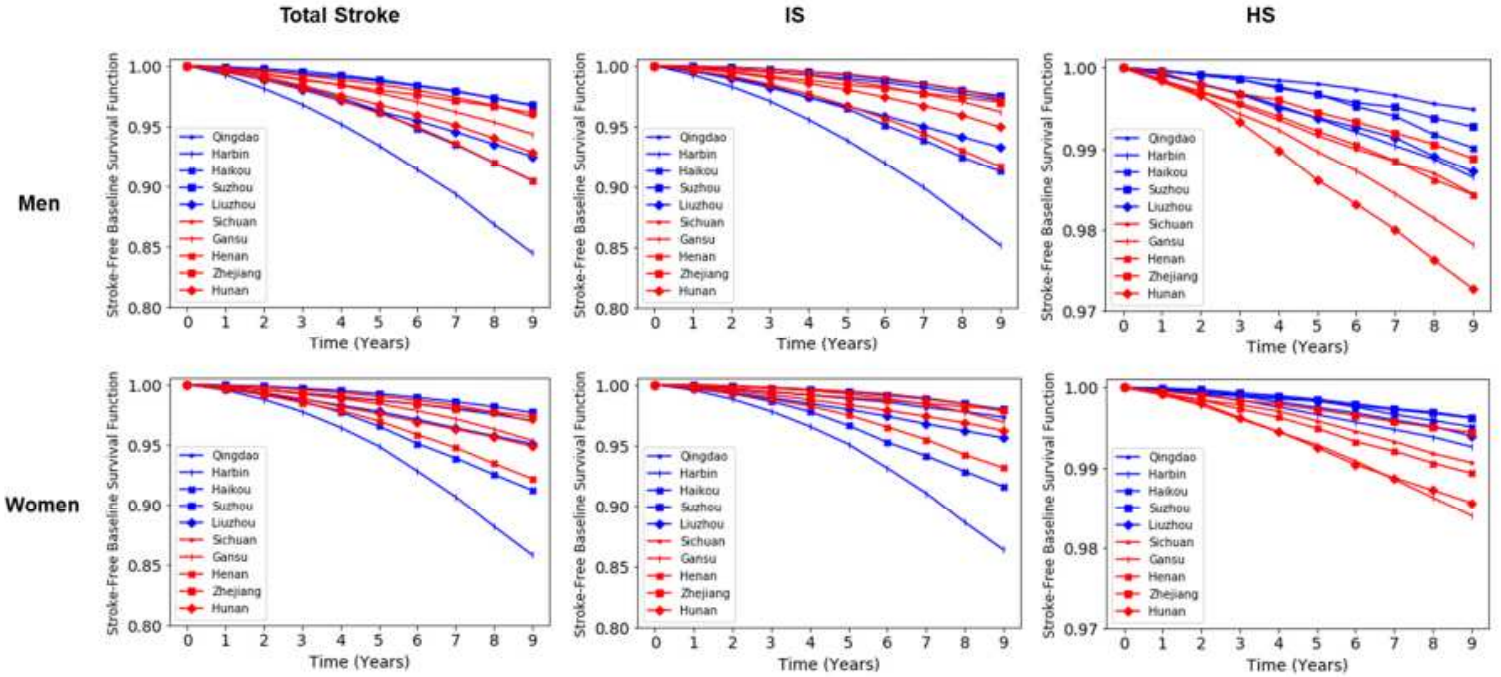
The worksheet titled "Model Params - Means" includes, for all CKB variables, the average risk factor values for the CKB men and women included in the present study. If any risk factor values are unknown for an individual, refer to this worksheet and enter the corresponding mean value into the "Stroke Risk Calculator" worksheet.

The worksheet titled "Calculation" pulls in the appropriate hazard ratios, mean risk factor values, individual risk factor values, and baseline survival function values corresponding to the user's input on the "Stroke Risk Calculator" worksheet. It then walks through the calculations performed by the selected model to generate a risk prediction, which is outputted in cell B16.

The worksheet titled "Risk Factor Definitions" includes the full list of CKB variables and their definitions.

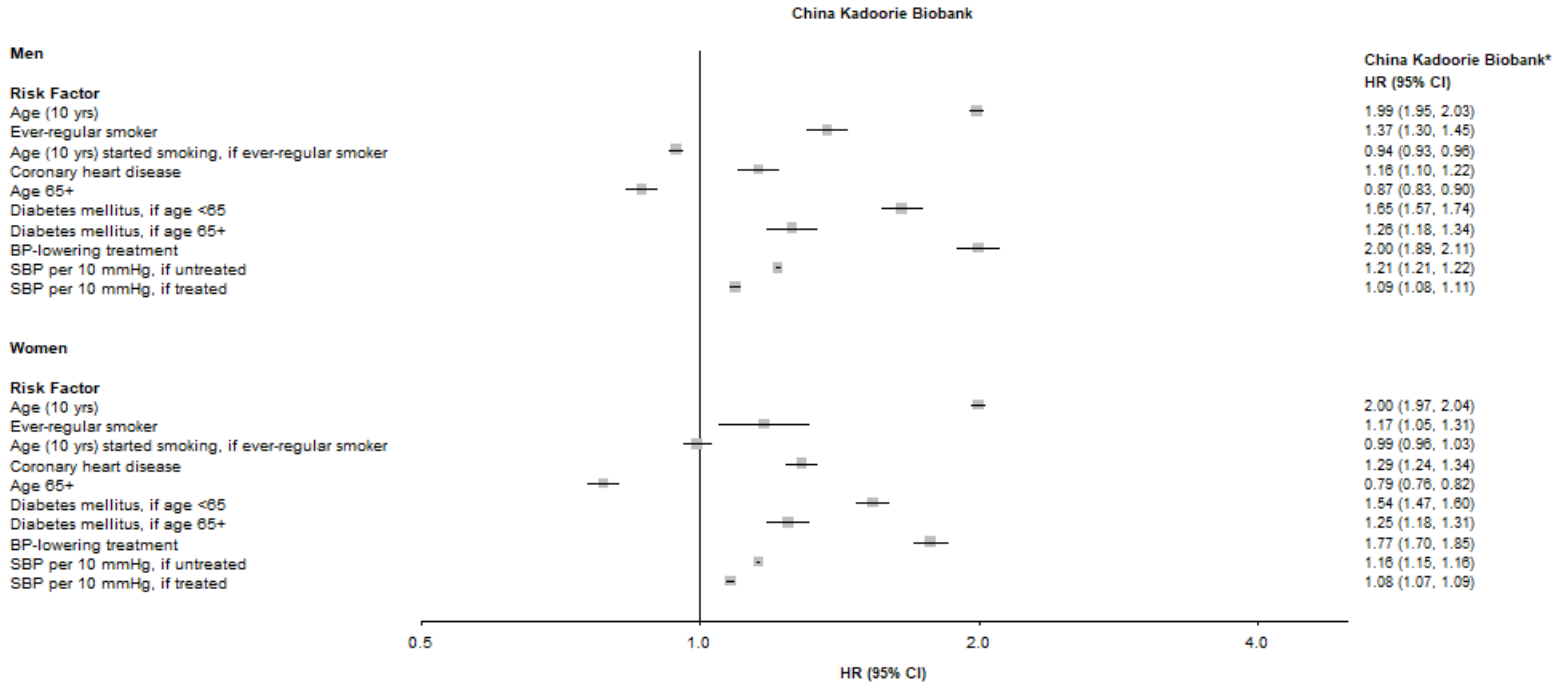
The worksheets titled "Risk Factor Qs" and "Options" are backend sheets used for displaying the appropriate risk factor questions and selection options, respectively, on the interactive "Stroke Risk Calculator" worksheet.

eFigure I. Area-specific and Aggregate CKB Baseline Survival Curves for First Total Stroke, First Ischemic Stroke, and First Hemorrhagic Stroke in Men and Women



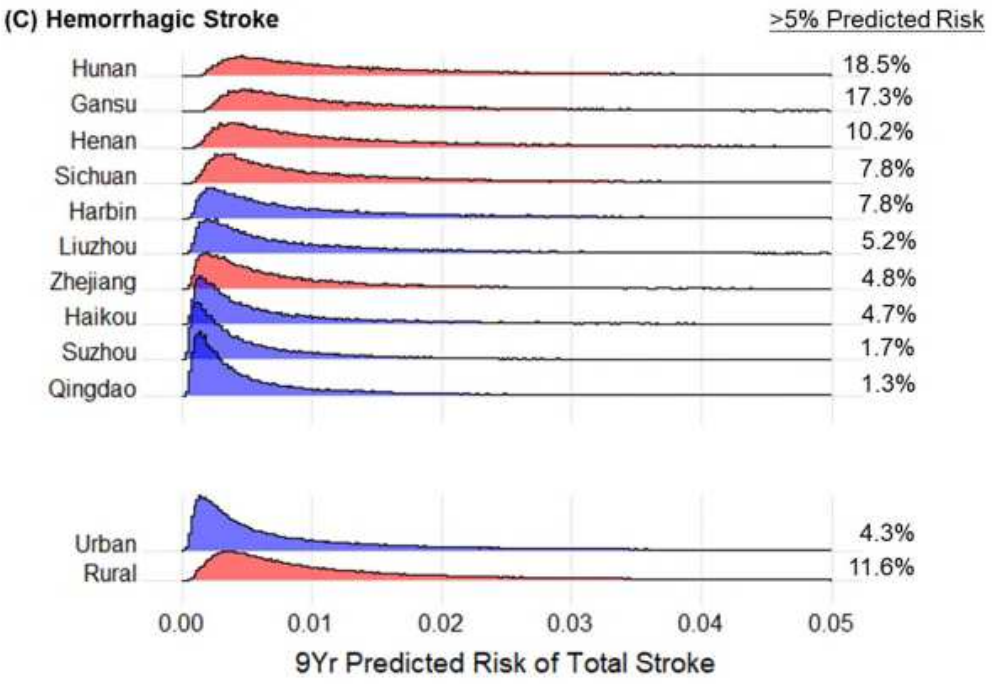
Note: Area-specific baseline survival functions are dependent on the risk factors included in the model as well as their corresponding hazard ratios. In this figure, the area-specific baseline survival functions are shown for recalibrated and refitted models with FSRP inputs only ("+Area stratification" models).

eFigure II. Multivariable Hazard Ratios (HR) and 95% CI for Total Stroke using Framingham ("2017 FSRP" Models) versus CKB ("+ Area Stratification" Models [Modified to Include Ever-Regular Smoking and Age Started Smoking]) Coefficients in Men and Women



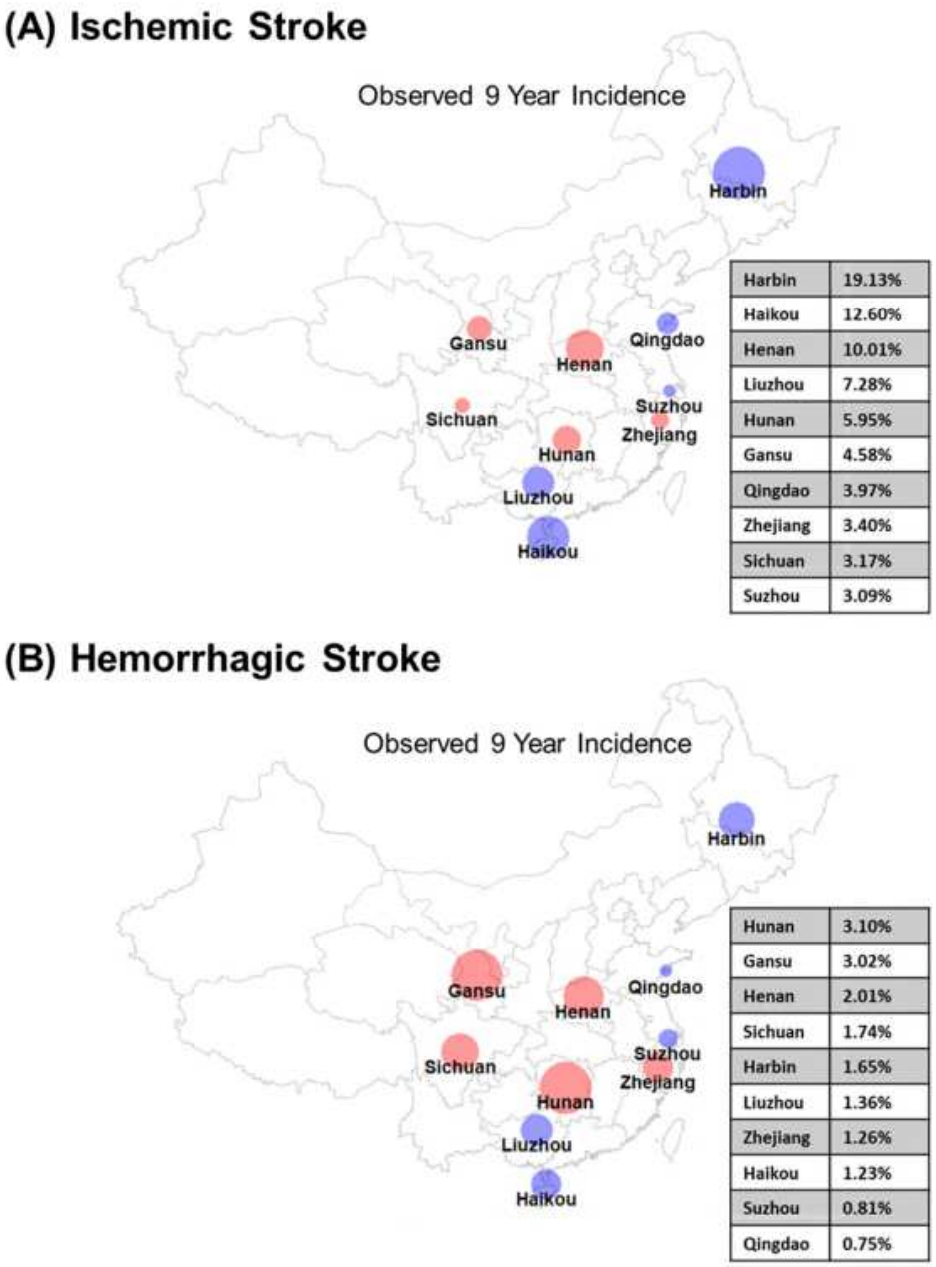
* The 2017 Framingham Stroke Risk Profile coefficients were refitted to China Kadoorie Biobank in a model including stratification by geographical area.

eFigure III. Predicted Risk Distributions for (A) Total Stroke, (B) Ischemic Stroke, and (C) Hemorrhagic Stroke by CKB Area.

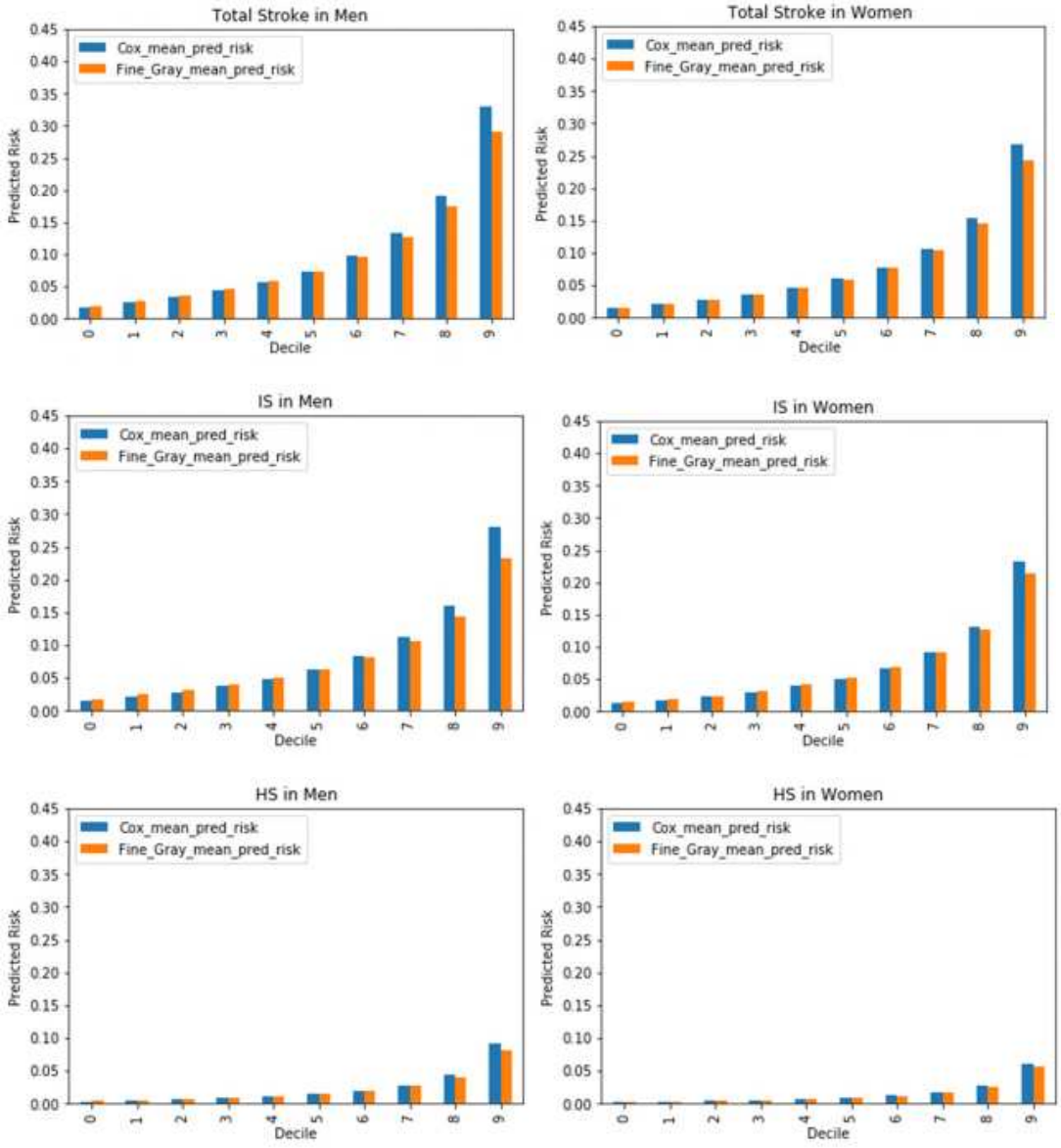


Note: Geographic areas in each subplot are ordered top-to-bottom by decreasing observed 9-year incidence. This corresponds well to the proportion of individuals identified as “high-risk” based on their predicted 9 year risk of stroke (>10% risk for total stroke and IS; >5% risk for HS).

eFigure IV. 9 Year Incidence for (A) Ischemic Stroke and (B) Hemorrhagic Stroke by CKB Area (Adjusted for Competing Risk of Death).

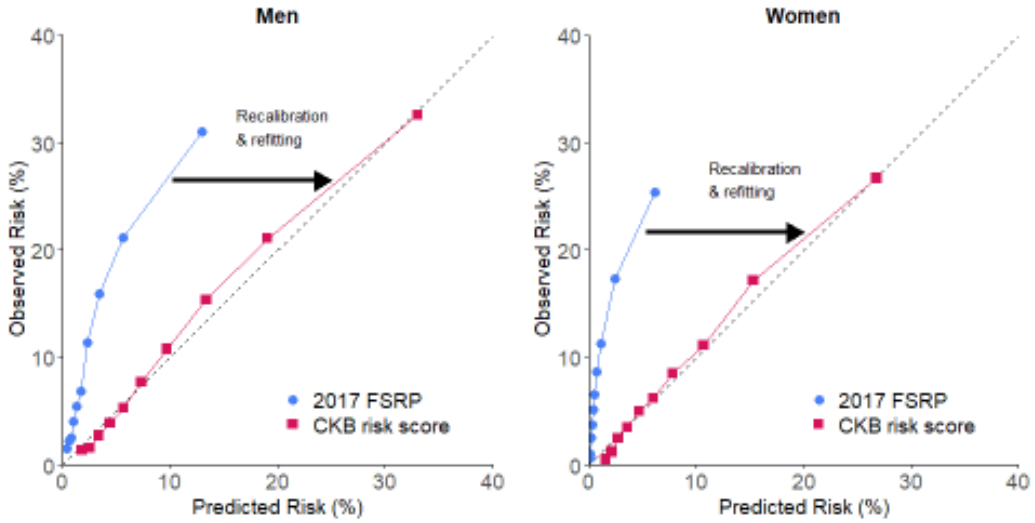


eFigure V. Comparison of 9-Year Risk Predictions in Test Set for Cox Model ("+Refitting") and Fine-Gray Model for Total Stroke, Ischemic Stroke, and Hemorrhagic Stroke in Men and Women.



Note: Both Cox and Fine-Gray models were specified with FSRP inputs. Hazard ratios for the Cox models and subdistribution hazard ratios for the Fine-Gray models are reported in eWorkbook 1.

eFigure VI. Calibration of the 2017 Framingham Stroke Risk Profile before and after recalibration and refitting in the China Kadoorie Biobank.



Calibration of the 2017 Framingham Stroke Risk Profile (FSRP) compared to a recalibrated and refitted risk score for prediction of total stroke in the China Kadoorie Biobank (CKB)

eTable I. Area-Specific Discrimination Performance for Total Stroke and Stroke Pathological Types in Men and Women Using "+ Area Stratification" Models with FSRP Inputs

Men

	Total Stroke AUC [95%CI]	**IS AUC [95%CI]**	**HS AUC [95%CI]**
Qingdao	0.84 [0.79-0.89]	0.85 [0.81-0.89]	0.81 [0.71-0.90]
Harbin	0.75 [0.73-0.77]	0.75 [0.73-0.77]	0.77 [0.73-0.81]
Haikou	0.80 [0.77-0.83]	0.80 [0.77-0.83]	0.79 [0.70-0.87]
Suzhou	0.82 [0.79-0.85]	0.82 [0.80-0.85]	0.72 [0.62-0.80]
Liuzhou	0.81 [0.79-0.83]	0.77 [0.75-0.80]	0.84 [0.79-0.89]
Sichuan	0.82 [0.79-0.85]	0.82 [0.78-0.86]	0.86 [0.83-0.89]
Gansu	0.80 [0.77-0.83]	0.75 [0.70-0.79]	0.86 [0.82-0.89]
Henan	0.78 [0.75-0.80]	0.77 [0.75-0.79]	0.74 [0.69-0.78]
Zhejiang	0.81 [0.78-0.84]	0.79 [0.75-0.83]	0.79 [0.74-0.83]
Hunan	0.80 [0.78-0.82]	0.78 [0.75-0.80]	0.81 [0.77-0.84]

Women

	Total Stroke AUC [95%CI]	**IS AUC [95%CI]**	**HS AUC [95%CI]**
Qingdao	0.81 [0.78-0.84]	0.79 [0.75-0.83]	0.78 [0.71-0.85]
Harbin	0.76 [0.74-0.77]	0.76 [0.74-0.77]	0.80 [0.76-0.84]
Haikou	0.80 [0.77-0.82]	0.77 [0.74-0.79]	0.85 [0.78-0.90]
Suzhou	0.84 [0.81-0.87]	0.82 [0.78-0.85]	0.74 [0.62-0.84]
Liuzhou	0.82 [0.79-0.84]	0.79 [0.77-0.82]	0.75 [0.69-0.81]
Sichuan	0.82 [0.79-0.84]	0.81 [0.78-0.84]	0.86 [0.82-0.90]
Gansu	0.79 [0.77-0.82]	0.77 [0.74-0.80]	0.86 [0.84-0.89]
Henan	0.77 [0.75-0.79]	0.77 [0.75-0.79]	0.77 [0.73-0.81]
Zhejiang	0.82 [0.79-0.85]	0.82 [0.80-0.85]	0.79 [0.73-0.85]
Hunan	0.76 [0.73-0.78]	0.76 [0.73-0.79]	0.79 [0.74-0.83]

eTable II. Area-Specific Discrimination Performance for Total Stroke and Stroke Pathological Types in Men and Women Using "+ Additional Risk Factors" Models

Men

	Total Stroke (66 Selected Risk Factors) **AUC [95%CI]**	**IS** (62 Selected Risk Factors) **AUC [95%CI]**	**HS** (42 Selected Risk Factors) **AUC [95%CI]**
Qingdao	0.84 [0.79-0.88]	0.86 [0.81-0.89]	0.84 [0.78-0.90]
Harbin	0.75 [0.73-0.77]	0.77 [0.75-0.79]	0.78 [0.73-0.83]
Haikou	0.79 [0.75-0.82]	0.81 [0.78-0.84]	0.79 [0.70-0.86]
Suzhou	0.83 [0.80-0.86]	0.83 [0.80-0.86]	0.77 [0.67-0.86]
Liuzhou	0.83 [0.80-0.85]	0.78 [0.76-0.81]	0.84 [0.80-0.88]
Sichuan	0.82 [0.79-0.85]	0.82 [0.78-0.86]	0.88 [0.84-0.91]
Gansu	0.81 [0.78-0.83]	0.77 [0.74-0.81]	0.85 [0.82-0.89]
Henan	0.79 [0.77-0.82]	0.77 [0.75-0.79]	0.77 [0.72-0.81]
Zhejiang	0.82 [0.79-0.85]	0.80 [0.76-0.83]	0.79 [0.73-0.84]
Hunan	0.81 [0.79-0.84]	0.79 [0.77-0.82]	0.83 [0.78-0.86]

Women

	Total Stroke (70 Selected Risk Factors) **AUC [95%CI]**	**IS** (80 Selected Risk Factors) **AUC [95%CI]**	**HS** (38 Selected Risk Factors) **AUC [95%CI]**
Qingdao	0.82 [0.79-0.85]	0.80 [0.76-0.83]	0.80 [0.73-0.86]
Harbin	0.76 [0.75-0.78]	0.76 [0.75-0.78]	0.81 [0.76-0.86]
Haikou	0.80 [0.78-0.82]	0.77 [0.75-0.79]	0.84 [0.77-0.90]
Suzhou	0.85 [0.82-0.87]	0.82 [0.78-0.86]	0.75 [0.63-0.85]
Liuzhou	0.82 [0.80-0.84]	0.80 [0.77-0.82]	0.74 [0.67-0.81]
Sichuan	0.82 [0.79-0.85]	0.82 [0.79-0.85]	0.86 [0.81-0.90]
Gansu	0.80 [0.77-0.83]	0.78 [0.75-0.81]	0.87 [0.84-0.90]
Henan	0.78 [0.77-0.80]	0.78 [0.76-0.80]	0.79 [0.74-0.83]
Zhejiang	0.83 [0.80-0.86]	0.83 [0.81-0.86]	0.81 [0.75-0.86]
Hunan	0.76 [0.74-0.79]	0.76 [0.74-0.79]	0.79 [0.74-0.83]

eTable III. Discrimination and Calibration Performance for Total Stroke and Stroke Pathological Types in Men and Women (Age 55-84) Using 2017 Framingham Stroke Risk Profile (FSRP)[7]

	Men		Women	
	Discrimination AUC [95% CI]	**Calibration χ^2**	**Discrimination AUC [95% CI]**	**Calibration χ^2**
Total Stroke				
2017 FSRP	0.68 [0.67-0.70]	1,346	0.66 [0.65-0.68]	2,021
+ Recalibration	0.68 [0.67-0.70]	235	0.66 [0.65-0.68]	248
+ Refitting	0.70 [0.69-0.71]	35	0.68 [0.67-0.69]	22
+ Area stratification	0.73 [0.72-0.74]	27	0.74 [0.73-0.75]	51
+ Additional risk factors	0.74 [0.73-0.75]	34	0.75 [0.74-0.76]	50
Ischemic Stroke (IS)				
2017 FSRP	0.68 [0.66-0.69]	928	0.65 [0.64-0.66]	1,588
+ Recalibration	0.68 [0.66-0.69]	209	0.65 [0.64-0.66]	274
+ Refitting	0.69 [0.68-0.71]	35	0.67 [0.66-0.68]	22
+ Area stratification	0.75 [0.74-0.76]	24	0.75 [0.74-0.76]	20
+ Additional risk factors	0.76 [0.74-0.77]	26	0.75 [0.74-0.76]	22
Hemorrhagic Stroke				
2017 FSRP	0.71 [0.69-0.74]	93	0.69 [0.67-0.71]	34
+ Recalibration	0.71 [0.69-0.74]	35	0.69 [0.67-0.71]	34
+ Refitting	0.72 [0.70-0.74]	15	0.71 [0.69-0.73]	12
+ Area stratification	0.75 [0.72-0.77]	5	0.75 [0.73-0.77]	28
+ Additional risk factors	0.76 [0.73-0.78]	3	0.77 [0.74-0.79]	11

Note: Atrial fibrillation was not recorded in the CKB.

eTable IV. Comparison of Discrimination and Calibration Performance of Different Risk Models for Prediction of Total Stroke and Stroke Pathological Types in Men and Women (With Modified Ordering of Incremental Changes Compared to Corresponding Table 2)

	Men		Women	
	Discrimination AUC [95% CI]	**Calibration χ^2**	**Discrimination AUC [95% CI]**	**Calibration χ^2**
Total Stroke				
2017 FSRP	0.78 [0.77-0.79]	1,825	0.77 [0.76-0.78]	3,053
+ Recalibration	0.78 [0.77-0.79]	156	0.77 [0.76-0.78]	506
+ Refitting	0.79 [0.79-0.80]	51	0.78 [0.77-0.78]	148
+ Additional risk factors	0.81 [0.80-0.81]	84	0.81 [0.80-0.81]	102
+ Area stratification	0.83 [0.82-0.84]	101	0.83 [0.82-0.84]	177
Ischemic Stroke (IS)				
2017 FSRP	0.77 [0.76-0.78]	1,200	0.76 [0.76-0.77]	2,406
+ Recalibration	0.77 [0.76-0.78]	118	0.76 [0.76-0.77]	479
+ Refitting	0.78 [0.78-0.79]	21	0.77 [0.76-0.78]	74
+ Additional risk factors	0.81 [0.80-0.81]	28	0.80 [0.79-0.81]	71
+ Area stratification	0.83 [0.82-0.84]	55	0.83 [0.82-0.84]	90
Hemorrhagic Stroke (HS)				
2017 FSRP	0.79 [0.78-0.81]	136	0.78 [0.76-0.80]	70
+ Recalibration	0.79 [0.78-0.81]	58	0.78 [0.76-0.80]	65
+ Refitting	0.80 [0.78-0.81]	23	0.80 [0.78-0.82]	33
+ Additional risk factors	0.82 [0.80-0.83]	32	0.82 [0.80-0.83]	6
+ Area stratification	0.82 [0.81-0.84]	14	0.82 [0.80-0.84]	9

Note: Shaded rows indicate changes relative to corresponding Table 2 due to including additional risk factors prior to area stratification.

References

1. Chen Y, Wright N, Guo Y, Turnbull I, Kartsonaki C, Yang L, et al. Mortality and recurrent vascular events after first incident stroke: a 9-year community-based study of 0.5 million Chinese adults. *Lancet Glob Health*. 2020;8:e580-90.
2. Davidson-Pilon C, Kalderstam J, Jacobson N, Reed S, Kuhn B, Zivich P, et al. (2019, April 26). CamDavidsonPilon/lifelines: v0.21.1 (Version v0.21.1). Zenodo. http://doi.org/10.5281/zenodo.2652543
3. Breslow NE. Discussion of the paper by D.R. Cox. *J R Statist Soc B*. 1972;34:215-16.
4. Lin DY. On the Breslow estimator. *Lifetime Data Anal*. 2007;13:471-80.
5. Friedman J, Hastie T, Tibshirani R. Glmnet: lasso and elastic-net regularized generalized linear models. R package version 3.0-2. 2019. http://CRAN.R-project.org/package=glmnet. Accessed April 8, 2020.
6. Hastie T, Qian J. Glmnet vignette. 2014. http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf. Accessed April 8, 2020.
7. Dufouil C, Beiser A, McLure LA, Wolf PA, Tzourio C, Howard VJ, et al. A revised Framingham Stroke Risk Profile to reflect temporal trends. *Circulation*. 2017;135:1145-1159.
8. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke*. 1991;22:312-18.
9. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke*. 1994;25:40-43.