BMJ Open

Development and internal validation of a multivariable prediction model for 6-year risk of stroke: a cohort study in middle-aged and elderly Chinese population

Qi Yu ,[1] Yuanzhe Wu,[1] Qingdong Jin,[2] Yanqing Chen,[1] Qingying Lin,[1] Xinru Liu[1]

To cite: Yu Q, Wu Y, Jin Q, *et al*. Development and internal validation of a multivariable prediction model for 6-year risk of stroke: a cohort study in middle-aged and elderly Chinese population. *BMJ Open* 2021;**11**:e048734. doi:10.1136/bmjopen-2021-048734

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-048734).

Received 08 January 2021
Accepted 10 June 2021

[1]Department of Scientific Research and Education, The First Hospital of Putian, Putian, Fujian, China
[2]Department of Neurosurgery, The First Hospital of Putian, Putian, Fujian, China

Correspondence to
Qi Yu; madyuqi@126.com

ABSTRACT

Objective To develop and internally validate a prediction model for 6-year risk of stroke and its primary subtypes in middle-aged and elderly Chinese population.

Design This is a retrospective cohort study from a prospectively collected database.

Participants We included a total 3124 adults aged 45–80 years, free of stroke or myocardial infarction at baseline in the 2009–2015 cohort of China Health and Nutrition Survey.

Primary and secondary outcome measures The outcome of the prediction model was stroke. Investigated predictors were: age, gender, body mass index (BMI), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), total cholesterol (TC), hypertension (HBP), drinking status, smoking status, diabetes and site. Stepwise multiple Cox regression was applied to identify independent predictors. A nomogram was constructed to predict 6-year risk of stroke based on the multiple analysis results. Bootstraps with 1000 resamples were applied to both C-index and calibration curve.

Result The overall incidence of overall stroke was 2.98%. Age, gender, HBP and TC were found as significant risk predictors for overall stroke; age, gender, HBP and LDL-C were found as significant risk predictors for ischaemic stroke; age, gender, HBP, BMI and HDL-C were found as significant risk predictors for haemorrhagic stroke. The nomogram was constructed using significant variables included in the model, with a C-index of 0.74 (95% CI: 0.72 to 0.76), 0.74 (95% CI: 0.71 to 0.77), and 0.81 (95% CI: 0.78 to 0.84) for overall stroke, ischaemic stroke, and haemorrhagic stroke model, respectively. The calibration curves demonstrated the good agreements between predicted and observed 6-year risk probability.

Conclusion Our nomogram could be convenient, easy to use and effective prognoses for predicting 6-year risk of stroke in middle-aged and elderly Chinese population.

Strengths and limitations of this study

- This is the first study to develop a nomogram for predicting 6-year incidence rate of total stroke and its subtypes in a Chinese population.
- Stepwise multiple regression and bootstrap internal validation were used to construct the models.
- The model performance was also assessed by imputing missing values using an imputational regression model.
- Because of lacking long-term follow-up data, the long-term result was uncertain.
- This study has a relatively high sample size when compared with previous studies on stroke. However, for the development of a prediction model, the sample size is relatively low, and therefore we had to conduct a strong variable selection and model validation procedure.

INTRODUCTION

Stroke is the leading cause of death and is a major cause of permanent disability worldwide.[1 2] In 1990–2010, the incidence of stroke had decreased by 12% in high-income countries, because of effective strategies for preventing cerebrovascular risk factors and good health services in developed countries. On the contrary, the age-adjusted incidence of stroke had significantly increased by 12% in low-income and middle-income countries.[3] In China, the world's most populous country, the incidence of stroke increased by 8.3% among adults aged 40 years and older from 2002 to 2013.[4 5]

Therefore, stroke prevention is essential for enhancing public health and reducing social burden in countries with heavy stroke burdens such as China. A prediction model for actively assessing stroke risk is required to ensure targeted strategies for stroke prevention and management in high-risk groups.

Previous studies that develop the stroke risk prediction models were initially developed from the western populations,[6–8] thus application of these models to the general Chinese population is still questionable. In addition, many stroke risk prediction models were only built for overall stroke.[9–11] However, there

are some notable differences in risk factors between ischaemic and haemorrhagic stroke.[12] Therefore, we developed a simple, convenient, and efficient model to predict the 6-year risk of overall, ischaemic and haemorrhagic stroke among middle-aged and elderly Chinese adults.

METHOD

Study design

The China Health and Nutrition Survey (CHNS) was conducted by the University of North Carolina at Chapel Hill and the National Institute for Nutrition and Health at the Chinese Center for Disease Control and Prevention in nine provinces (Heilongjiang, Liaoning, Shandong, Jiangsu, Henan, Hubei, Hunan, Guangxi and Guizhou) to examine the economic status, public resources, health and nutrition. A multistage, random cluster process was used to draw the samples in the Chinese population. The survey was an ongoing nationwide study that started in 1989 and subsequently conducted in 1991, 1993, 1997, 2000, 2004, 2006, 2009, 2011 and 2015. All participants provided written informed consent. Details about the study design are available elsewhere.[13 14] Baseline data collection included demographic information, medical history, standardised medical examination, laboratory tests and anthropometric measurements.

Study population

For this study, data were drawn from the 2009–2015 CHNS cycles (n=12 178), after excluding participants who were younger than 45 years or older than 80 years at baseline (n=6148); persons who had a history of stroke or myocardial infarction (n=291); persons who lost to follow-up (n=1770); and those without complete physical survey data or blood measure data at baseline (n=845). As a result, 1434 men and 1690 women were available for analysis.

Data collection

History of diseases, individual activities, lifestyle, health status, marriage and birth history were acquired through an individual questionnaire. Adults and children received detailed physical examinations that include weight, height, arm and head circumference, mid-arm skinfold measurements and blood pressure. Blood pressure was measured thrice by experienced physicians with the participant in the sitting position. The biomarker data collected in CHNS 2009 involve the release of 26 fasting blood measures on individuals aged 7 years and older. Frozen serum samples were sent to a national central lab in Beijing for measurement of serum lipid levels.

Definitions

Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were defined as mean SBP and DBP of three test results. Hypertension (HBP) was defined as blood pressure >160/90 mm Hg or taking antihypertensive drugs. Smoking status was classified into three categories as follows: never smoker, ever smoker and current smoker; alcohol drinking status was divided into two groups: never drinker and ever or current drinker. Diabetes was identified by self-reports of a history of diabetes diagnosis.

Incident stroke was defined by a doctor's diagnosis or treatment history for stroke during the follow-up period (2009–2015). Cases were censored at the date of diagnosis of stroke or the final visit, whichever came first.

Statistical analysis

In this analysis, we included the following clinical candidate predictors: age (year), gender (male or female), body mass index (BMI, kg/m^2), low-density lipoprotein cholesterol (LDL-C, mmol/L), high-density lipoprotein cholesterol (HDL-C, mmol/L), total cholesterol (TC, mmol/L), HBP (no or yes), drinking status (no or ever/current), smoking status (no, ever or current), diabetes (no or yes), and site (urban or rural). Results were presented as mean±SD for continuous variables, and number (percentage) for categorical variables. To determine statistical significance, analysis of variance and chi-square tests were used for continuous variables and categorical variables according to sex-specific groups.

Cox proportional hazards models were used to explore the relationship between baseline risk factors and stroke incidence. In addition, we also constructed the models with the outcomes of ischaemic stroke and haemorrhagic stroke (participants who had an unknown stroke were excluded). The proportional hazards assumption of the Cox models was examined using Schoenfeld residuals. The significance of each risk factor in the cohort was assessed by univariate Cox regression analysis for investigating the independent variable of incident stroke. Risk factors associated with stroke at a significant level were further evaluated using multivariate Cox proportional hazards regression analysis by the stepwise selection method ($p<0.1$). The independent predictors related to incident stroke were used to build a nomogram. A score was assigned to each risk factor in the nomogram so that total points could be easily calculated to estimate the probability of stroke. To assess the accuracy of the nomogram, bootstraps with 1000 resamples were applied to both C-index and calibration curve. Moreover, we did stratified analyses to explore whether the association of risk factors with overall stroke varied across gender. We also assessed the models by imputing missing values using an imputational regression model (with the transcan function in R).

Internal validation is the process of determining internal validity. Bootstrap is an important internal verification method. Bootstrap samples are drawn with replacement from the original sample, reflecting the drawing of samples from an underlying population. Bootstrap samples are of the same size as the original sample. For bootstrap validation, a prediction model is developed in each bootstrap sample. This model is evaluated both in the bootstrap sample and in the original sample. The bootstrap is used to estimate the optimism: the

decrease between performance in the bootstrap sample and performance in the original sample. This optimism is subsequently subtracted from the original estimate to obtain an optimism-corrected performance estimate.[15 16]

We can use the bootstrap approach to any performance measure, including the C-index and calibration measures. The C-index is the most commonly used performance measure to depict the discriminative ability of generalised linear regression models. For a binary outcome, C-index is identical to the area under the receiver operating characteristic (ROC) curve. The ROC curve is a plot of the sensitivity (true-positive rate) versus 1-specificity (false-positive rate) for continuous cut-offs for the probability of an outcome. The area under the curve can be interpreted as the probability that a patient with the outcome is given a higher probability of the outcome by the model than a randomly chosen patient without the outcome.[17] Another important performance of a model is calibration: the consistency between observed results and predictions. For example, if we predict 70% probability of stroke event for a person, the observed frequency of stroke event should be 70 out of 100 such persons. Calibration curves describe the calibration of each model in terms of the agreement between the predicted risks of stroke and observed frequency of stroke. The y-axis represents the actual stroke rate. The x-axis represents the predicted stroke risk. The grey line represents a perfect prediction by an ideal model. The red solid line represents the performance of the nomogram, of which a closer fit to the grey line indicates a better prediction.

All statistical analyses were carried out using R software V.3.6.1 (http://www.R-project.org).

RESULT

Baseline clinical characteristics of participants

The baseline clinical characteristics of the patients are listed in table 1. During a mean follow-up of 5.89 years, 2.98% (93) participants developed first ever stroke events, including 37 ischaemic strokes, 25 haemorrhagic strokes and 31 unspecified stroke events.

Multivariable analysis predicting stroke

Variables identified as predictors of incident stroke are listed in table 2. Significant risk factors in univariate analysis (P<0.2) were included in the stepwise multiple Cox regression models. The proportional hazards assumption of the Cox models was examined for three prediction models (overall stroke, ischaemic stroke and haemorrhagic stroke) using Schoenfeld residuals, which showed no significant departure from proportionality over time (p>0.05).

Table 3 displays the HR (95% CI) for stroke in different subgroups. Stepwise multiple Cox regression analysis showed that age (for each 1-year increase, HR (95% CI)=1.07 (1.04 to 1.09)), gender (for female vs male, HR (95% CI)=1.66 (1.09 to 2.52)), HBP (for no vs yes, HR (95% CI)=2.63 (1.73 to 3.98)) and TC (for each 1 mmol/L increase, HR (95% CI)=1.26 (1.03 to 1.53)) were independent risk predictors for developing overall stroke, which were further used to build a nomogram. The significant risk factors for ischaemic stroke were age (for each 1-year increase, HR (95% CI)=1.07 (1.03 to 1.11)), gender (for female vs male, HR (95% CI)=1.85 (0.96 to 3.58)), HBP (for no vs yes, HR (95% CI)=2.29 (1.19 to 4.40)) and LDL-C (for each 1 mmol/L increase, HR (95% CI)=1.27 (1.07 to 1.51)). The significant risk factors for haemorrhagic stroke were age (for each 1-year increase, HR (95% CI)=1.08 (1.03 to 1.13)), gender (for female vs male, HR (95% CI)=3.49 (1.43 to 8.55)), HBP (for no vs yes, HR (95% CI)=2.72 (1.18 to 6.27)), BMI (for each 1 kg/m^2 increase, HR (95% CI)=1.13 (1.02 to 1.25)) and HDL-C (for each 1 mmol/L increase, HR (95% CI)=1.58 (1.19 to 2.09)).

In stratified analysis, HBP, elevated HDL-C level, ever smoking, and ever or current drinking were associated with an increased risk of overall stroke for women than for men. However, in sensitivity analyses using imputed dataset, HDL-C was not identified to be associated with overall stroke in different gender (online supplemental material 1, Figures S1 and S2).

Nomogram construction and validation

These independently associated risk factors were used to construct an overall stroke, ischaemic stroke and haemorrhagic stroke risk estimation nomogram (figures 1 and 2). Moreover, nomograms were also built using imputed dataset (online supplemental material 1, Figures S3 and S4). A score was assigned on the point scale for each subtype within these variables. By counting the total score and locating it on the total points scale, we could easily draw a straight line down to determine the estimated probability of stroke over 6 years.

The resulting model was internally validated using the bootstrap validation method. The nomogram demonstrated good accuracy in estimating the risk of overall stroke, ischaemic stroke and haemorrhagic stroke, with an unadjusted C-index of 0.74 (95% CI: 0.72 to 0.76), 0.74 (95% CI: 0.71 to 0.77) and 0.81 (95% CI: 0.78 to 0.84), respectively; a bootstrap-corrected C-index of 0.73, 0.72 and 0.78, respectively. The calibration plots for 6-year overall stroke-free, ischaemic stroke-free and haemorrhagic stroke-free probability showed good agreement between the predicted possibility and the actual observation (figures 3 and 4).

We also assessed the models by imputing missing values using an imputational regression model. Before data imputation, there were 3969 cases. HBP was missing in 28 cases (0.7%); BMI was missing in 152 cases (3%); LDL-C was missing in 397 (10%) cases; HDL-C was missing in 397 (10%) cases; TC was missing in 397 (10%) cases; smoking status was missing in 26 (0.7%) cases; drinking status was missing in 27 (0.7%) cases; diabetes was missing in 19 (0.5%) cases. Finally, the missing values were imputed using a single conditional imputation method (with the transcan function in R). Our findings remained robust

Table 1 Baseline clinical characteristics of participants

Characteristics	Total	Gender		P value
		Female	Male	
Age (year)	58.80±8.64	58.83±8.72	58.75±8.54	0.880
Gender				
Female	1690 (54.10%)	NA	NA	
Male	1434 (45.90%)	NA	NA	
BMI (kg/m^2)	23.72±3.43	23.94±3.53	23.47±3.30	<0.001
HBP				0.075
No	2283 (73.08%)	1257 (74.38%)	1026 (71.55%)	
Yes	841 (26.92%)	433 (25.62%)	408 (28.45%)	
LDL-C (mmol/L)	3.14±1.00	3.24±0.98	3.02±1.01	<0.001
HDL-C (mmol/L)	1.46±0.45	1.48±0.44	1.43±0.45	<0.001
TC (mmol/L)	5.05±0.98	5.15±0.99	4.94±0.96	<0.001
Smoking status				<0.001
No	2123 (67.96%)	1628 (96.33%)	495 (34.52%)	
Ever	106 (3.39%)	6 (0.36%)	100 (6.97%)	
Current	895 (28.65%)	56 (3.31%)	839 (58.51%)	
Drinking status				<0.001
No	2058 (65.88%)	1535 (90.83%)	523 (36.47%)	
Ever/current	1066 (34.12%)	155 (9.17%)	911 (63.53%)	
Diabetes				0.009
No	2713 (86.84%)	1475 (87.28%)	1238 (86.33%)	
Yes	411 (13.16%)	215 (12.72%)	196 (13.67%)	
Site				0.992
Urban	906 (29.00%)	490 (28.99%)	416 (29.01%)	
Rural	2218 (71.00%)	1200 (71.01%)	1018 (70.99%)	
Stroke type				0.072
Ischaemic stroke	37 (39.78%)	15 (37.50%)	22 (41.51%)	
Haemorrhagic stroke	25 (26.88%)	7 (17.50%)	18 (33.96%)	
Unknown	31 (33.33%)	18 (45.00%)	13 (24.53%)	

BMI, body mass index; HBP, hypertension; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; NA, not applicable; TC, total cholesterol.

with imputed data, with an unadjusted C-index of 0.73 (95% CI: 0.70 to 0.76), 0.70 (95% CI: 0.68 to 0.72), and 0.83 (95% CI: 0.81 to 0.85) for overall stroke, ischaemic stroke and haemorrhagic stroke, respectively; a bootstrap-corrected C-index of 0.72, 0.68 and 0.81 for overall stroke, ischaemic stroke and haemorrhagic stroke, respectively. Moreover, the calibration plots for 6-year overall stroke-free, ischaemic stroke-free and haemorrhagic stroke-free probability showed good agreement between the predicted possibility and the actual observation (online supplemental material 1, Figures S5 and S6). These results indicated that the nomogram could accurately predict 6-year risk of overall stroke, ischaemic stroke and haemorrhagic stroke in middle-aged and elderly Chinese population.

DISCUSSION

We used data from the CHNS Study, which is a nationally representative prospective cohort in nine provinces around China, enabling us to provide a general method to estimate the risk of stroke in middle-aged and elderly Chinese population. We created a practical nomogram based on age, gender, HBP and TC to predict the risk of overall stroke; a nomogram based on age, gender, HBP and LDL-C to predict the risk of ischaemic stroke; a nomogram based on age, gender, HBP, BMI and HDL-C to predict the risk of haemorrhagic stroke (figures 1 and 2). Discrimination was supported by the unadjusted C-index of 0.74 (95% CI: 0.72 to 0.76), 0.74 (95% CI: 0.71 to 0.77) and 0.81 (95% CI: 0.78 to 0.84), respectively; bootstrap-corrected C-index of 0.73,

Table 2 Univariate Cox regression analysis of stroke incidence

Covariate	Overall stroke		Ischaemic stroke		Haemorrhagic stroke	
	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value
Age(year)	1.08 (1.05 to 1.10)	<.001	1.08 (1.04 to 1.12)	<.001	1.08 (1.04 to 1.13)	0.001
Gender, female versus male	1.57 (1.04 to 2.36)	0.032	1.73 (0.90 to 3.34]	0.101	3.04 (1.27 to 7.27)	0.013
HBP, no versus yes	3.35 (2.23 to 5.05)	<.001	2.94 (1.54 to 5.60)	0.001	4.18 (1.88 to 9.29)	<.001
BMI (kg/m^2)	1.05 (0.99 to 1.11)	0.075	1.03 (0.94 to 1.12)	0.589	1.13 (1.02 to 1.25)	0.022
LDL-C (mmol/L)	1.22 (1.05 to 1.41)	0.009	1.31 (1.09 to 1.57]	0.003	1.08 (0.75 to 1.55)	0.679
HDL-C (mmol/L)	1.24 (0.90 to 1.71)	0.187	1.14 (0.62 to 2.09)	0.672	1.54 (1.12 to 2.13)	0.008
TC (mmol/L)	1.28 (1.06 to 1.56)	0.011	1.47 (1.09 to 1.97)	0.010	1.12 (0.76 to 1.65)	0.572
Smoking status, no versus ever	2.74 (1.31 to 5.72)	0.007	2.48 (0.75 to 8.22)	0.137	2.76 (0.63 to 12.07)	0.178
Smoking status, no versus current	0.99 (0.62 to 1.57)	0.950	0.85 (0.40 to 1.82)	0.679	1.26 (0.53 to 2.97)	0.596
Drinking status, no versus ever/current	1.22 (0.80 to 1.85)	0.350	1.32 (0.68 to 2.54)	0.413	1.52 (0.69 to 3.34)	0.302
Diabetes, no versus yes	1.48 (0.88 to 2.51)	0.141	2.14 (1.01 to 4.53)	0.048	0.91 (0.27 to 3.03)	0.873
Site, urban versus rural	0.95 (0.61 to 1.47)	0.804	0.85 (0.43 to 1.69)	0.644	1.29 (0.52 to 3.24)	0.584

BMI, body mass index; HBP, hypertension; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol.

0.72 and 0.78, respectively. In addition, among the three different models we established, the haemorrhagic stroke model showed the best discrimination predictive ability (unadjusted C-index=0.81). Calibration curves for 6-year overall, ischaemic and haemorrhagic stroke-free probability demonstrated the good agreements between predicted and observed probability (figures 3 and 4). In addition, the sensitivity analysis using imputed data did not change the results substantially (online supplemental material 1, Figures S3–S6).

Many stroke risk prediction models aiming to predict the risk of stroke have been developed previously, for example, the new FSRP,[6] the QStroke[7] and others.[8] The new FSRP represented the current status of stroke predictors in the USA and France[6] ; however, its performance among Chinese residents has not been evaluated. In China, Wang *et al*[9] developed an age-specific and sex-specific lifetime risk chart of stroke including six risk factors (blood pressure, non-HDL-C, HDL-C, BMI, diabetes and smoking), which allowed for stratification of lifetime risk of stroke. In addition, Chien *et al*[11] developed a model from a community cohort of 3513 Chinese participants aged ≥35 years old in Taiwan incorporating six risk factors including age, gender, BP, family history of stroke, atrial fibrillation and diabetes (C-index of 0.77). However, these risk prediction models are only available for the total stroke risk. Rare stroke risk assessment tool is available for either ischaemic stroke and haemorrhagic stroke in China. Therefore, compared with current existing models, we think that better understanding of

Table 3 Stepwise multiple Cox regression analysis of stroke incidence (p<0.1)

Covariate	Overall stroke		Ischaemic stroke		Haemorrhagic stroke	
	HR (95% CI)	P value	HR (95% CI)	P value	HR (95% CI)	P value
Age (year)	1.07 (1.04 to 1.09)	<0.001	1.07 (1.03 to 1.11)	<0.001	1.08 (1.03 to 1.13)	<0.001
Gender, female versus male	1.66 (1.09 to 2.52)	0.018	1.85 (0.96 to 3.58)	0.068	3.49 (1.43 to 8.55)	0.006
HBP, no versus yes	2.63 (1.73 to 3.98)	<0.001	2.29 (1.19 to 4.40)	0.013	2.72 (1.18 to 6.27)	0.019
BMI (kg/m^2)	NA	NA	NA	NA	1.13 (1.02 to 1.25)	0.021
LDL-C (mmol/L)	NA	NA	1.27 (1.07 to 1.51)	0.006	NA	NA
HDL-C (mmol/L)	NA	NA	NA	NA	1.58 (1.19 to 2.09)	0.002
TC (mmol/L)	1.26 (1.03 to 1.53)	0.024	NA	NA	NA	NA

BMI, body mass index; HBP, hypertension; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; NA, not applicable; TC, total cholesterol.

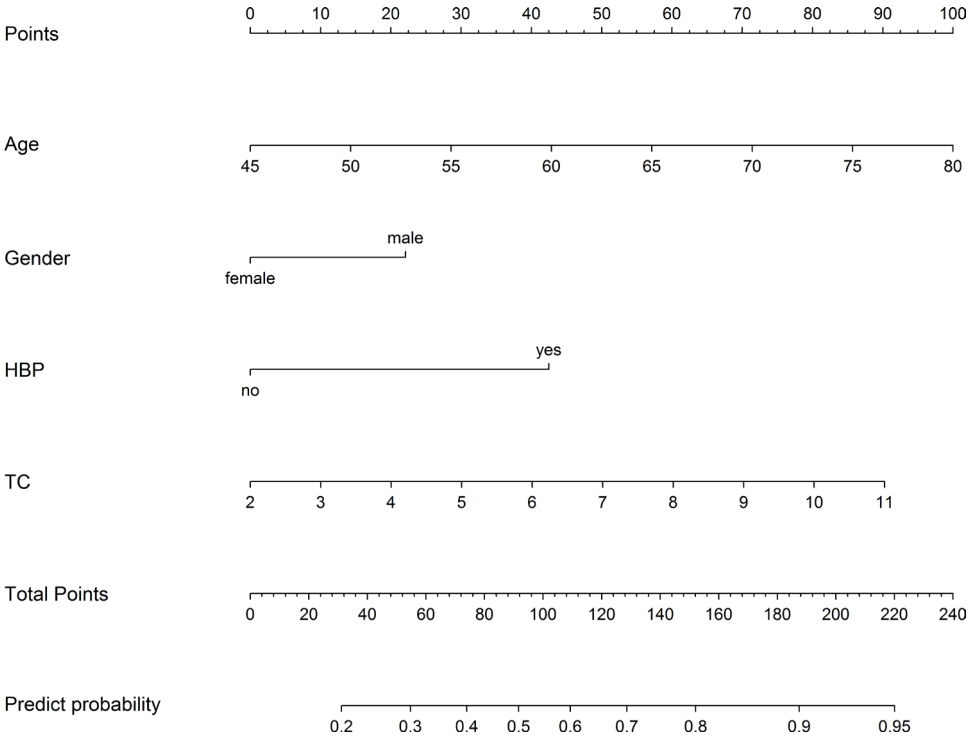

Figure 1 Nomogram for predicting 6-year risk of overall stroke for middle-aged and elderly Chinese population. Measurement: age (year), BMI (kg/m^2), LDL-C (mmol/L), HDL-C (mmol/L) and TC (mmol/L). The scores corresponding to each factor are listed on the 'Points' axis. To estimate the 6-year probability of disease, calculate the sum of the scores of each factor and locate the sum on the 'Total Points' axis, then read the probability on the 'Predict probability' axis. BMI, body mass index; HBP, hypertension; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol.

the association of risk factor with subtype-specific stroke and stroke model derived from the CHNS population would be more enlightening for stroke prevention among Chinese adults.

HBP is an important and modifiable predictive risk factor for stroke.[18 19] Results from the INTERSTROKE Study show that HBP, using a definition: history of HBP or blood pressure >160/90 mm Hg, was a highly

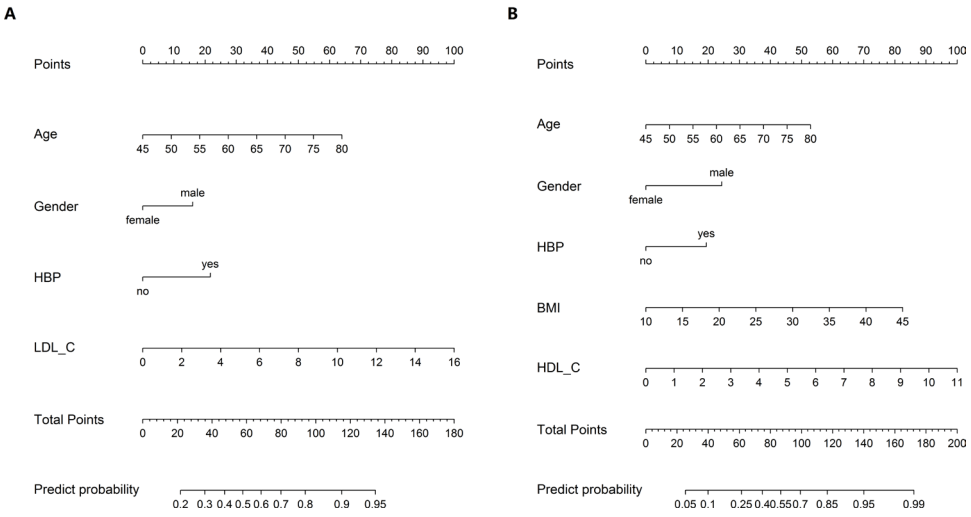

Figure 2 Nomogram for predicting 6-year risk of stroke for middle-aged and elderly Chinese population: (A) ischaemic stroke, (B) haemorrhagic stroke. Measurement: age (year), BMI (kg/m^2), LDL-C (mmol/L), HDL-C (mmol/L) and TC (mmol/L). The scores corresponding to each factor are listed on the 'Points' axis. To estimate the 6-year probability of disease, calculate the sum of the scores of each factor and locate the sum on the 'Total Points' axis, then read the probability on the 'Predict probability' axis. BMI, body mass index; HBP, hypertension; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol.

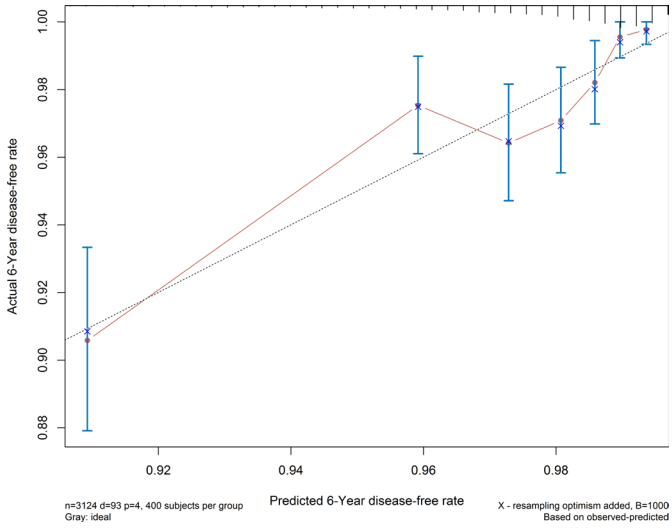

Figure 3 Calibration curves for the nomogram (overall stroke). Nomogram-predicted probability and observed frequency over 6 years for stroke among participants were plotted in the x-axis and y-axis, respectively. The grey line indicates the ideal plot for the calibration curve, where the nomogram predicted probabilities perfectly match the observed probabilities in all subgroups.

important risk factor for stroke in developing countries and accounts for 35% of all stroke.[20] Our result confirmed the association between HBP and risk of ischaemic and haemorrhagic stroke, and a more potent association for haemorrhagic stroke than for ischaemic stroke (table 3). In China, the prevalence of HBP rapidly increased in the past 30 years. However, awareness, treatment and control of HBP declined or remained unchanged in China from 2000 to 2010,[4 21] while they had increased significantly in the developed countries,[22] thus these may affect the incidence of stroke.

The association between cholesterol level and stroke is complex, with an increased risk of ischaemic stroke with elevated TC,[23–25] and a decreased risk of haemorrhagic stroke with elevated TC.[26 27] LDL-C showed positive associations with ischaemic stroke, and no association with haemorrhagic stroke.[23] HDL-C showed positive associations with haemorrhagic stroke in a recent meta-analysis,[28] but not by a recent retrospective cohort study.[23] In our study (table 3), the association with the risk of overall stroke, ischaemic stroke and haemorrhagic stroke was positive for TC, LDL-C and HDL-C, respectively. Further, our analysis indicated that BMI exerted an impact on the onset of haemorrhagic stroke. A notable correlation between BMI and haemorrhagic stroke has also been reported,[29 30] although conflicting studies exist.[31 32] Moreover, our result found that the risk of stroke was higher for ever smokers than for current smokers without adjustment for covariates (table 2). However, our research lacked detailed data about quantity or intensity of smoking, and the proportion of heavy smokers among the ever smokers might be higher than current smokers in our cohort.

We also found that HBP, elevated HDL-C level, ever smoking, and ever or current drinking were more strongly associated with overall stroke in women than for men. However, in sensitivity analysis with imputed data, HDL-C was not identified to be associated or showed only weak relationships with different gender (online supplemental material 1, Figures S1 and S2). These results were similar to a previous observational study.[33]

A nomogram is an excellent visual tool that is convenient, easy to use and effective prognoses, which can easily estimate the risk of stroke by adding the score of each predicting variable to predicted probability. In addition, early treatment can be given to patients at high risk

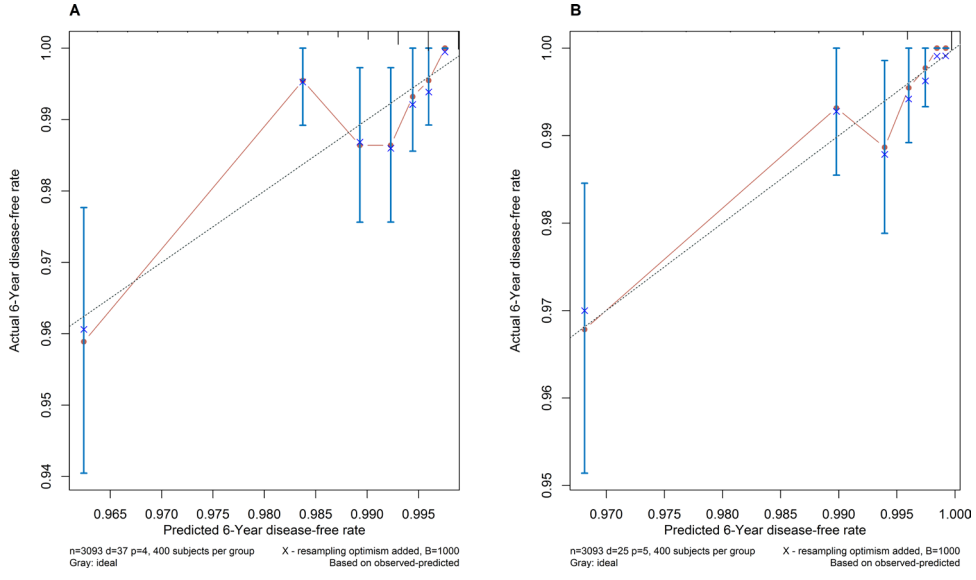

Figure 4 Calibration curves for the nomogram ((A) ischaemic stroke, (B) haemorrhagic stroke). Nomogram-predicted probability and observed frequency over 6 years for stroke among participants were plotted in the x-axis and y-axis, respectively. The grey line indicates the ideal plot for the calibration curve, where the nomogram predicted probabilities perfectly match the observed probabilities in all subgroups.

of early stroke. In contrast, those patients at low risk of stroke can be managed expectantly to prevent the potential treatment. Therefore, patients and doctors can use the nomogram to assist the level 1 prevention of stroke and provide key information to reduce the incidence and burden of diseases. Furthermore, non-professionals can learn about the nomogram through the internet, TV or newspaper in a short period of time, thereby improving their understanding of the risk factors of diseases.

Our study has several limitations. First, for the development of a prediction model, the sample size is relatively low. The prediction of risk factors for stroke in China still requires more data to improve the prediction model. Second, the long-term, >6 years, incidence of stroke for these risk factors is unknown. In future studies, we project to prospectively follow up participants, gather data about mortality and incidence of stroke in participants. Finally, the nomogram has not been externally validated. However, we used a 1000-bootstrap resampling strategy for internal validation and the nomogram showed good performance in terms of calibration and discrimination for predicting risk of stroke in middle-aged and elderly Chinese population.

CONCLUSION

In summary, the nomogram developed here can be conveniently used to facilitate the individualised prediction of 6-year risk of stroke in middle-aged and elderly Chinese population.

Acknowledgements This research uses data from CHNS. We thank the National Institute of Nutrition and Food Safety, China Centre for Disease Control and Prevention, Carolina Population Centre, the University of North Carolina at Chapel Hill, and the Fogarty International Centre, National Institute of Health for financial support for the CHNS data collection and analysis files.

Contributors QY conceptualised and designed the study, carried out the initial analyses, drafted the initial manuscript, and reviewed and revised the manuscript. YW, QJ, YC, QL and XL critically reviewed and revised the manuscript. All authors approved the final manuscript for submission.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Data of this study were from the China Health and Nutrition Survey (CHNS). The Institutional Review Committees of the North Carolina at Chapel Hill, and the National Institute for Nutrition and Health at the Chinese Center for Disease Control and Prevention approved the survey protocols and instruments and the process for obtaining informed consent for the survey. All participants (or their parents or guardians) agreed to participate in the survey and provided written informed consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data from China Health and Nutrition Survey was used in this study, which can be downloaded at http://www.cpc.unc.edu/projects/china/data/datasets.

ORCID iD

Qi Yu http://orcid.org/0000-0002-1138-8368

REFERENCES

1 Feigin VL, Norrving B, Mensah GA. Global burden of stroke. Circ Res 2017;120:439–48.
2 Mozaffarian D, Benjamin EJ, *et al*, Writing Group Members. Heart disease and stroke statistics-2016 update: a report from the american heart association [published correction appears in circulation. 2016 Apr 12;133(15):e599]. Circulation 2016;133:e38–60.
3 Feigin VL, Forouzanfar MH, Krishnamurthi R. Global and regional burden of stroke during 1990-2010: findings from the global burden of disease study 2010 [published correction appears in lancet. 2014 Jan 18;383(9913):218]. Lancet 2014;383:245–54.
4 Guan T, Ma J, Li M, *et al*. Rapid transitions in the epidemiology of stroke and its risk factors in China from 2002 to 2013. Neurology 2017;89:53–61.
5 Wang Z, Hu S, Sang S, *et al*. Age-period-cohort analysis of stroke mortality in china: data from the global burden of disease study 2013. Stroke 2017;48:271–5.
6 Dufouil C, Beiser A, McLure LA, *et al*. Revised framingham stroke risk profile to reflect temporal trends. Circulation 2017;135:1145–59.
7 Hippisley-Cox J, Coupland C, Brindle P. 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. BMJ 2013;346:f2573.
8 Chambless LE, Heiss G, Shahar E. Prediction of ischemic stroke risk in the atherosclerosis risk in communities study [published correction appears in Am J Epidemiol. 2004 Nov 1;160(9):927]. Am J Epidemiol 2004;160:259–69.
9 Wang Y, Liu J, Wang W, *et al*. Lifetime risk of stroke in young-aged and middle-aged Chinese population: the Chinese Multi-Provincial cohort study. J Hypertens 2016;34:2434–40.
10 Xing X, Yang X, Liu F, *et al*. Predicting 10-year and lifetime stroke risk in Chinese population. Stroke 2019;50:2371–8.
11 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.
12 Boehme AK, Esenwa C, Elkind MSV. Stroke risk factors, genetics, and prevention. Circ Res 2017;120:472–95.
13 Popkin BM, Du S, Zhai F, *et al*. Cohort Profile: the china health and nutrition survey--monitoring and understanding socio-economic and health change in china, 1989-2011. Int J Epidemiol 2010;39:1435–40.
14 Yan S, Li J, Li S, *et al*. The expanding burden of cardiometabolic risk in china: the china health and nutrition survey. Obes Rev 2012;13:810–21.
15 Efron B, Tibshirani RJ. *An introduction to the bootstrap*. CRC press, 1994.
16 Harrell FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
17 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
18 Chobanian AV, Bakris GL, Black HR, *et al*. Seventh report of the joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension 2003;42:1206–52.
19 Lewington S, Clarke R, Qizilbash N. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies [published correction appears in Lancet. 2003 Mar 22;361(9362):1060]. Lancet 2002;360:1903–13.
20 O'Donnell MJ, Xavier D, Liu L, *et al*. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the interstroke study): a case-control study. Lancet 2010;376:112–23.
21 Wu S, Wu B, Liu M, *et al*. Stroke in China: advances and challenges in epidemiology, prevention, and management. Lancet Neurol 2019;18:394–405.

22 Mills KT, Bundy JD, Kelly TN, *et al*. Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. Circulation 2016;134:441–50.
23 Gu X, Li Y, Chen S, *et al*. Association of lipids with ischemic and hemorrhagic stroke: a prospective cohort study among 267 500 Chinese. Stroke 2019;50:3376–84.
24 Kurth T, Everett BM, Buring JE, *et al*. Lipid levels and the risk of ischemic stroke in women. Neurology 2007;68:556–62.
25 Wannamethee SG, Shaper AG, Ebrahim S. Hdl-cholesterol, total cholesterol, and the risk of stroke in middle-aged British men. Stroke 2000;31:1882–8.
26 Iso H, Jacobs DR, Wentworth D, *et al*. Serum cholesterol levels and six-year mortality from stroke in 350,977 men screened for the multiple risk factor intervention trial. N Engl J Med 1989;320:904–10.
27 Zhang X, Patel A, Horibe H. Asia Pacific cohort studies collaboration. cholesterol, coronary heart disease, and stroke in the Asia Pacific region. Int J Epidemiol 2003;32:563–72.
28 Wang X, Dong Y, Qi X, *et al*. Cholesterol levels and risk of hemorrhagic stroke: a systematic review and meta-analysis. Stroke 2013;44:1833–9.
29 Bazzano LA, Gu D, Whelton MR, *et al*. Body mass index and risk of stroke among chinese men and women. Ann Neurol 2010;67:11–20.
30 Song Y-M, Sung J, Davey Smith G, *et al*. Body mass index and ischemic and hemorrhagic stroke: a prospective study in korean men. Stroke 2004;35:831–6.
31 Yonemoto K, Doi Y, Hata J, *et al*. Body mass index and stroke incidence in a japanese community: the Hisayama study. Hypertens Res 2011;34:274–9.
32 Strazzullo P, D'Elia L, Cairella G, *et al*. Excess body weight and incidence of stroke: meta-analysis of prospective studies with 2 million participants. Stroke 2010;41:e418–26.
33 Peters SAE, Carcel C, Millett ERC, *et al*. Sex differences in the association between major risk factors and the risk of stroke in the UK biobank cohort study. Neurology 2020;95:e2715–26.