Development of Model to Predict 10-Year Risk of Ischemic and Hemorrhagic Stroke and Ischemic Heart Disease Using the China Kadoorie Biobank

Author(s):

Songchun Yang, PhD candidate; Yuting Han, PhD candidate; Canqing Yu, PhD; Yu Guo, MSc; Yuanjie Pang, PhD;
Dianjianyi Sun, PhD; Pei Pei, MSc; Ling Yang, PhD; Yiping Chen, DPhil; Huaidong Du, PhD; Hao Wang, MSc; M Sofia Massa, PhD; Derrick Bennett, PhD; Robert Clarke, FRCP; Junshi Chen, MD; Zhengming Chen, DPhil; Jun Lv, PhD; Liming Li, MPH on behalf of the China Kadoorie Biobank Collaborative Group

This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (CC BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Neurology® Published Ahead of Print articles have been peer reviewed and accepted for publication. This manuscript will be published in its final form after copyediting, page composition, and review of proofs.

Errors that could affect the content may be corrected during these processes.
Corresponding Author:
Jun Lv, lvjun@bjmu.edu.cn

Affiliation Information for All Authors: 1. Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing 100191, China; 2. Peking University Center for Public Health and Epidemic Preparedness & Response, Beijing 100191, China; 3. Fuwai Hospital Chinese Academy of Medical Sciences, Beijing 100730, China; 4. Chinese Academy of Medical Sciences, Beijing 100730, China; 5. Medical Research Council Population Health Research Unit at the University of Oxford, Oxford OX3 7LF, United Kingdom; 6. Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, United Kingdom; 7. NCDs Prevention and Control Department, Zhejiang CDC, Hangzhou 310051, China; 8. China National Center for Food Safety Risk Assessment, Beijing 100738, China; 9. Key Laboratory of Molecular Cardiovascular Sciences (Peking University), Ministry of Education, Beijing 100191, China

Equal Author Contribution: 
Jun Lv and Liming Li contributed equally to this work

Contributions:
Songchun Yang: Drafting/revision of the manuscript for content, including medical writing for content; Analysis or interpretation of data
Yuting Han: Analysis or interpretation of data
Canqing Yu: Major role in the acquisition of data; Analysis or interpretation of data

Copyright © 2022 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology.
Yu Guo: Major role in the acquisition of data
Yuanjie Pang: Analysis or interpretation of data
Dianjianyi Sun: Analysis or interpretation of data
Pei Pei: Major role in the acquisition of data
Ling Yang: Major role in the acquisition of data
Yiping Chen: Major role in the acquisition of data
Huaidong Du: Major role in the acquisition of data
Hao Wang: Major role in the acquisition of data
M Sofia Massa: Drafting/revision of the manuscript for content, including medical writing for content
Derrick Bennett: Drafting/revision of the manuscript for content, including medical writing for content
Robert Clarke: Drafting/revision of the manuscript for content, including medical writing for content
Junshi Chen: Study concept or design
Zhengming Chen: Study concept or design
Jun Lv: Drafting/revision of the manuscript for content, including medical writing for content; Study concept or design
Liming Li: Major role in the acquisition of data; Study concept or design; Other

Figure Count:
4

Table Count:
2

Search Terms:
[ 6 ] Infarction, Chinese population, hemorrhagic stroke, ischemic heart disease, risk prediction
Acknowledgment:

The most important acknowledgment is to the participants in the study and the members of the survey teams in each of the 10 regional centers, as well as to the project development and management teams based at Beijing, Oxford and the 10 regional centers. The members of the China Kadoorie Biobank collaborative group are listed in Appendix 2.

Study Funding:

This work was supported by National Natural Science Foundation of China (82192904, 82192901, 82192900). The CKB baseline survey and the first re-survey were supported by a grant from the Kadoorie Charitable Foundation in Hong Kong. The long-term follow-up is supported by grants from the UK Wellcome Trust (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z), grants (2016YFC0900500) from the National Key R&D Program of China, National Natural Science Foundation of China (81390540), and Chinese Ministry of Science and Technology (2011BAI09B01). Derrick Bennett was supported by the National Institute of Health Research (NIHR) Biomedical Research Center Oxford. Zhengming Chen was supported by the UK Medical Research Council.

Disclosures:

The authors report no disclosures relevant to the manuscript.

Handling Editor Statement:

Abstract

Background and Objectives: Contemporary cardiovascular disease (CVD) risk prediction models are rarely applied in routine clinical practice in China due to substantial regional differences in absolute risks of major CVD types within China. Moreover, the inclusion of blood lipids in most risk prediction models also limits their use in the Chinese population. We developed 10-year CVD risk prediction models excluding blood lipids that may be applicable to diverse regions of China.
Methods: We derived sex-specific models separately for ischemic heart disease (IHD), ischemic stroke (IS), and hemorrhagic stroke (HS) in addition to total CVD in the China Kadoorie Biobank. Participants were aged 30–79 years without CVD at baseline. Predictors included age, systolic and diastolic blood pressure, use of blood pressure-lowering treatment, current daily smoker, diabetes, and waist circumference. Total CVD risks were combined in terms of conditional probability using the predicted risks of three submodels. Risk models were recalibrated in each region by two methods ("practical" and "ideal") and risk prediction was estimated before and after recalibration.

Results: Model derivation involved 489,596 individuals, including 45,947 IHD, 43,647 IS, and 11,168 HS cases during 11 years of follow-up. In women, the Harrell's C was 0.732 (95% CI 0.706-0.758), 0.759 (0.738-0.779), and 0.803 (0.778-0.827) for IHD, IS, and HS, respectively. The Harrell's C for total CVD was 0.734 (0.732-0.736), 0.754 (0.752-0.756), and 0.774 (0.772-0.776) for models before recalibration, after practical recalibration, and after ideal recalibration. The calibration performances improved after recalibration, with models after ideal recalibration showing the best model performances. The results for men were comparable to those for women.

Discussion: Our CVD risk prediction models yielded good discrimination of IHD and stroke subtypes in addition to total CVD without including blood lipids. Flexible recalibration of our models for different regions could enable more widespread use using resident health records covering the overall Chinese population.

Classification of Evidence: This study provides Class I evidence that a prediction model incorporating accessible clinical variables predicts 10-year risk of ischemic heart disease, ischemic stroke, and hemorrhagic stroke in the Chinese population, age 30–79 years.

Abbreviations and acronyms

CI: confidence interval
Introduction

Cardiovascular diseases (CVD), including ischemic heart disease (IHD), ischemic stroke (IS), and hemorrhagic stroke (HS), are the leading causes of the global burden of diseases.\(^1\) Risk prediction models are important tools for identifying high-risk individuals for early interventions in the primary prevention of CVD. Although the availability of many models has prompted researchers to focus on tailoring and improving existing models in local populations,\(^2\) there is currently no model that simultaneously has the following three characteristics and can be widely used in Chinese population.

The first is to distinguish the risks between IS and HS. The Chinese population has
a much higher incidence of HS than Western populations.\textsuperscript{1} However, currently recommended models worldwide, including China, either did not include HS as the outcome\textsuperscript{3, 4} or did not distinguish IS from HS.\textsuperscript{5-9} The second is to achieve efficient and widespread use of the risk prediction model by linking to dynamic electronic health records.\textsuperscript{3, 6, 10-12} In China, establishing resident health records (RHRs) has been one of the basic services provided by the National Basic Public Health Service program (NBPHSP) since 2009, covering the whole population in the Chinese mainland. The free RHRs for the whole population include sociodemographic information, personal and family medical history, lifestyle, and non-invasive physical examinations (e.g., height, weight, waist circumference [WC], and blood pressure). Free blood cholesterol measurements are only provided to people \( \geq 65 \) years. However, previous models mostly rely on blood lipids.\textsuperscript{3-7, 9}

The third is to be widely applicable to different regions. Risk factors for different CVD subcategories may vary or have different effect estimates.\textsuperscript{13-15} For a model that predicts the risk of composite outcome (single-model estimated, such as total CVD), the magnitude of association (i.e., regression coefficient in the model) might be affected by the composition proportion of each CVD subcategory in the derivation dataset, leading to potentially biased estimations. Therefore, considering that the absolute risks of three CVD subcategories vary substantially among different regions in China,\textsuperscript{16} previous risk prediction models based on single-model approach are less applicable to broad areas of China.\textsuperscript{3-7, 9} To address this issue, the World Health Organization (WHO) CVD Risk Chart Working Group has proposed a new modeling and recalibration approach to help adapt the original models to different regions.\textsuperscript{8}

Based on the China Kadoorie Biobank (CKB), one of the largest population-based cohorts globally, the present report primarily aimed to develop pragmatic disease-specific and overall CVD risk prediction models based on widely available variables in RHRs of China, covering IHD, IS, and HS. Next, we evaluated two model recalibration approaches based on the method proposed by the WHO CVD
Risk Chart Working Group to adapt our models to different regions. We further evaluated the potential incremental value of more predictors to the derived models.

Methods

Study Population

CKB is an ongoing prospective study of 512,725 participants aged 30 to 79 who were enrolled from five urban and five rural regions situated in the northeast, northwest, east, south, and southwest of China during 2004–2008. The study regions were selected according to local disease patterns, exposure to risk factors, population stability, quality of death and disease registries, and local commitment and capacity. Within each study region, all potentially eligible participants in each of 100-150 rural villages or urban residential committees were invited to take part in the survey. The estimated population response rate was about 30% (26–38% in the five rural regions and 16–50% in the five urban regions). Details of the study have been described elsewhere. Briefly, all participants had valid baseline data, including a complete interviewer-administered laptop-based questionnaire and physical measurements conducted by trained health workers using calibrated instruments and standard protocols. A 10-ml random blood sample was collected for each participant with the time of the last meal recorded.

Study Design

In this study, four interrelated components were involved (
Figures

Figure 1). (1) Predictor selection. We selected several predictors from a predefined list of candidate variables for model derivation. (2) Model derivation. We then derived sex-specific and outcome-specific 10-year CVD risk prediction models among participants without CVD at baseline, namely, the "CKB-CVD models". (3) Demonstration of model recalibration. To support the flexible updating of the derived models, we compared two methods of recalibration. (4) Evaluation of model updating. Based on the derived models, we further evaluated the predictive utility of CVD risk factors that were not included in the model derivation process.

Assessment of candidate predictors

We predefined 16 candidate predictors (eTable 1 in the supplement) for model derivation according to the following criteria: (1) established or probable risk factors of CVD; (2) widely available in RHRs; and (3) available in the CKB baseline survey. Other CVD risk factors that are not widely available in RHRs (eTable 1) were used to explore the ability to improving the derived models.

The baseline questionnaire collected information on sociodemographic characteristics, lifestyle behaviours, dietary habits, personal and family medical history. All participants were asked, "How often do you smoke tobacco now?" Four response options were: (1) do not smoke now, (2) only occasionally, (3) on most days, and (4) every day. Current daily smokers were defined as those who answered (4). Systolic blood pressure (SBP) and diastolic blood pressure (DBP) at baseline were measured using a UA-779 digital sphygmomanometer. Measurements were made
after a minimum of five minutes of sitting and repeated twice. If the two measurements for SBP were more than 10 mmHg apart, a third measurement was taken. Only the last two readings were recorded and averaged for further analyses. Use of blood pressure-lowering treatment was self-reported by participants and was defined as answering "yes" to questions "Is your diagnosed hypertension still on treatment?" or "Did you take any drugs to lower blood pressure in the last two days?". History of diabetes was defined as self-reported diabetes diagnosed by a physician before the baseline survey. We did not use random blood glucose measurements as our previous study did, because of its limited availability in China.\textsuperscript{18} WC was measured midway between the iliac crest and the lower rib margin at the end of normal expiration using a flexible plastic tape to the nearest 0.1 cm. The assessment and definition of other candidate predictors were shown in eTable 1 (supplement).

**Ascertainment of Outcomes**

All participants were followed up for incident disease outcomes since their enrolment at baseline. Incident events were identified by using linkages with local disease and death registries and the national health insurance system, and supplemented by active follow-up.\textsuperscript{17} The loss to follow-up was <1\% prior to censoring on December 31, 2017. Trained staff blinded to baseline information coded all events using International Classification of Diseases, Tenth Revision (ICD-10). In this study, IHD events included fatal or nonfatal angina (I20), myocardial infarction (I21–I23), and other IHD (I24–I25); IS events included fatal or nonfatal cerebral infarction (I63); HS events included fatal or nonfatal subarachnoid (SAH) or intracerebral hemorrhage (ICH) (I60–I62). Unspecified stroke events were coded as I64. The total CVD was defined as fatal or nonfatal IHD and stroke (I20–I25, I60–I64). Since 2014, medical records of incident IHD and stroke cases have been retrieved and reviewed by qualified cardiovascular specialists blinded to baseline exposures of patients. By October 2018, of the retrieved medical records of 33,515 IHD cases, 34,758 IS cases, and 5023 HS cases, the confirmed rates of the diagnosis were 87.9\%, 91.5\%, and 80.4\% for IHD, IS,
Statistical analysis

Cox proportional hazard models were used to develop prediction models, stratified by ten study regions and with time-on-study as the time scale. Time-on-study was the time from the baseline to the first of the following: first diagnosis of CVD, death, loss to follow-up, or December 31, 2017. All analyses were performed separately for women and men. Discrimination performance was accessed by Harrell's C statistic. Calibration performance was assessed graphically by comparing the mean predicted risks at 10 years with the observed risks across deciles of predicted risks. The Nam-D'Agostino test was used to quantify the agreement or fit.

In the predictor selection component, the basic model included four well-established CVD risk factors as predictors: baseline age (years), SBP (mmHg), current daily smoker (yes or no), and history of diabetes (yes or no). Interactions between age and the other three predictors were also included. Based on the relative integrated discrimination improvement (IDI) in the total CVD risk prediction model, three additional predictors, "use of blood pressure-lowering treatment" (yes or no), DBP (mmHg), and WC (cm) were also selected for model derivation (eMethods, eTable 2, eTable 3 in the supplement).

In the model derivation component, we developed models separately for IHD, IS, and HS ("submodel" of the "CKB-CVD models", and HS, separately. 
Figures

Figure 1) to allow separate recalibration to the disease-specific incidences in different regions. Predictors included the aforementioned seven variables and interactions between age and the other six variables. Before deriving the models, the necessity of logarithmic transformation of all continuous variables was examined, since the natural logarithm is the most common way to handle continuous variables in previous risk prediction models.\(^5,7\) Finally, all continuous variables were modelled as simple linear form and were centred (age at 55 years, SBP at 120 mmHg, DBP at 80 mmHg, and waist circumference at 80 cm) to provide a more straightforward interpretation of the regression estimates and facilitate recalibration of the models.\(^3,6,8\)

The proportional hazards assumption was assessed by plotting the scaled Schoenfeld residuals versus time. Either non-existent or minimal deviations were observed. Baseline survival estimate at 10 years \(S_0(10)\) was estimated by pooling the \(S_0(10)\) across regions weighted by the number of cases that had occurred by 10 years.\(^8\) To avoid overfitting, we used an internal-external cross-validation approach (IECV) in which each region was left out of the model derivation and used to calculate a validation Harrell's C in turn.\(^22\) The total CVD risk was combined in terms of conditional probability using the predicted risks calculated by the three submodels ("combined model"), i.e., \(Risk_{CVD} = 1 - (1 - Risk_{IHD}) \times (1 - Risk_{IHS}) \times (1 - Risk_{HS})\).\(^8\) We checked the assumption of independence among risks of the three CVD subcategories by considering competing risks.\(^8\)

To simulate the process of applying the "CKB-CVD models" in different regions, model recalibration was performed separately for ten study regions. We calculated the...
ideal performance metrics assuming the availability of observed 10-year risks in each study region (“ideal recalibration”). However, in practice, the observed 10-year risks of the target population are difficult to obtain. The recalibration process could also use age-specific (5-year age groups) and sex-specific mean risk factor levels and annual incidence estimates of CVD in each study region (“practical recalibration”) (eMethods in the supplement). Model performance before and after recalibration was evaluated in the whole derivation dataset and each study region. We also evaluated model performance before and after recalibration in participants <65 and ≥65 years, with and without hypertension, and with and without diabetes.

We further explored whether the remaining predictors (eTable 1 in the supplement) could be used for updating the "CKB-CVD models" in the future. We added each predictor individually and evaluated their relative IDI for each major CVD subtype, as we did in the predictor selection component outlined above.

This study adhered to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement for reporting. Analyses were done with Stata 15.0. Figures were produced using R 3.6.0.

**Standard Protocol Approvals, Registrations, and Patient Consents**

CKB had ethical approvals from the Ethical Review Committee of the Chinese Center for Disease Control and Prevention (Beijing, China) and the Oxford Tropical Research Ethics Committee, University of Oxford (UK). All participants provided a written informed consent form.

**Data availability**

Cohort description and questionnaires are available at ckbiobank.org/site/. Statistical code is available from Dr. Lv (e-mail, lvjun@bjmu.edu.cn). Details of how to access China Kadoorie Biobank data and details of the data release schedule are available from ckbiobank.org/site/Data+Access.
Results

Overview of the study population

The study population included 302,522 women and 210,203 men aged 30–79 years at the baseline survey. Participants who had been diagnosed with IHD (n=15,472) or stroke (n=7657) before the baseline survey were excluded from the derivation dataset (Figure 1). No participant had missing data on the main predictors used in the model derivation. The remaining 489,596 participants had a median age of 52 (p25–p75: 43–60) years for women and 50 (42–58) years for men (Table 1). During a median of 11 years of follow-up, 86,464 (17.7%) participants had CVD events, including 45,947 IHD, 43,647 IS, and 11,168 HS events. The standardised 10-year risks of three CVD subcategories varied greatly across regions (eFigure 1 in the supplement).

The newly-derived "CKB-CVD models"

Seven predictors were used to derive the sex-specific "CKB-CVD models" separately for three CVD subcategories (Table 2). The adjusted hazard ratios (HRs) of predictors varied across different sexes and different CVD subcategories. For example, HRs of DBP (per 10 mmHg) ranged from 1.05 (95% CI: 1.04–1.07) for IHD in women to 1.33 (1.29–1.38) for HS in men; HRs of WC (per 10 cm) ranged from 0.92 (0.89–0.95) for HS in women to 1.16 (1.14–1.17) for IHD in women. Regression coefficients for
each predictor are provided in eAppendix 1. An example risk calculation is shown in eMethods (supplement). The assumption of independence among risks of three CVD subcategories was not violated (eFigure 2 in the supplement). The HS submodel had the best discrimination performance in internal validation (Figure 2). Calibration performances of the three submodels were poor in most study regions (eFigure 3 in the supplement). In sensitivity analysis, we derived models separately for ICH, whose beta coefficients were similar to those of HS models (eTable 4 in the supplement).

As for the "combined model" before recalibration ("original model" in the following), Harrell's C was 0.734 (95% CI: 0.732–0.736) for women and 0.743 (0.741–0.746) for men in the derivation dataset (Figure 3), ranging from 0.705 (0.701–0.710) for women in Harbin to 0.804 (0.795–0.813) for women in Suzhou (eFigure 4 in the supplement). Calibration was poor in the derivation dataset (Figure 3) and most study regions (eFigure 5).

Recalibration of the "CKB-CVD models"

After "practical recalibration", Harrell's C increased to 0.754 (95% CI: 0.752–0.756) for women and 0.764 (0.762–0.766) for men, and calibration improved for both sexes in the derivation dataset. Models after "ideal recalibration" showed the best model performances (Figure 3). Harrell's C in each study region almost remained unchanged after "practical recalibration" and "ideal recalibration" (data not shown). Calibration performances improved in almost all study regions after "ideal recalibration" and in half of the study regions after "practical recalibration". The number of urban regions with improved calibration performance after "practical recalibration" was greater than those in rural regions (eFigure 5 in the supplement). An example of practical recalibration is shown in eMethods. We also provided an interactive 10-year CVD risk calculator and a calculator of practical recalibration parameters in eAppendix 1.

Discrimination of the "original model" was lower in the older (≥65 years), hypertension patients, and diabetes patients, especially in older women (Harrell's C=0.584, 95% CI: 0.579–0.589) (eFigure 6). Separate recalibration in each study
region improved the discrimination and the calibration among these participants (eFigure 7-9).

Evaluation of model updating
Based on the CKB-CVD models, only "level of education" and "waist-hip ratio" in men and "total physical activity" in women had relative IDI greater than 1% for the HS submodel. Other predictors had a small or non-effect on improving the three submodels (Figure 4).

Classification of Evidence
This study provides Class I evidence that a prediction model incorporating accessible clinical variables predicts 10-year risk of ischemic heart disease, ischemic stroke, and hemorrhagic stroke in the Chinese population, age 30-79 years.

Discussion
In this large population-based Chinese cohort, we developed pragmatic sex-specific risk models that predict 10-year risks of CVD subcategories and overall CVD. Given the high burden of HS in the Chinese population, our models distinguish stroke subtypes and should be more applicable to the Chinese than previous models. Our models achieved good discrimination of risk groups, even without using blood lipids information, indicating a potentially wider use based on RHRs. Flexible model updating methods improved the model performance when used in different regions.

Two latest CVD primary prevention guidelines for Chinese have recommended two 10-year CVD risk prediction models as risk assessment tools, one derived from the Chinese Multi-provincial Cohort Study (CMCS) and the other derived from the China-PAR Project.\textsuperscript{24, 25} The loss to follow-up in both cohorts, 21.3%\textsuperscript{26} and 9.9%,\textsuperscript{7} respectively, may increase the risk of bias according to PROBAST.\textsuperscript{27} Also, both models relied on blood lipids, used "hard" IHD as the outcome, and did not consider the risk of HS separately.\textsuperscript{4, 7} Few studies have derived risk prediction models for HS. One study was performed in a small cohort of 4400 Chinese steelworkers recruited...
before 1980, with only 33 HS events in the derivation dataset. The other study was conducted using three population-based cohorts from the United States and the Netherlands and included 325 intracerebral hemorrhage (ICH) events. As for IHD, previous studies have reported that patients with angina also have a poor prognosis. The simultaneous risk prediction of "hard" IHD and "soft" stroke might lead to misunderstandings among model users. Therefore, consistent with QRISK3 and PREDICT, we included such outcomes in high-risk population screening.

Most previous models only included SBP as the predictor. In the present study, the addition of DBP to the basic model for HS substantially improved the predictive ability. Although the relative IDI for the IHD and IS submodels was small, the category-free net reclassification improvement was > 10% for IS and >5% for IHD in both sexes (data not shown). Body mass index was included as a predictor in the WHO non-laboratory-based model. However, it did not improve the predictive ability as much as WC in our analysis (the predictor selection component), consistent with the finding of the China-PAR Project. We, therefore, included DBP and WC as predictors, given their incremental values in outcome prediction and wide availability in RHRs of China. Our models used the self-reported history of diabetes as a predictor, consistent with the practice of RHRs. Despite nearly half of the undiagnosed diabetes being excluded in our cohort, the addition of screen-detected diabetes only led to a minimal improvement of the three submodels.

Older individuals and individuals with hypertension or diabetes are major users of health services in China. We examined model performances stratified by these characteristics and found models had the lowest discrimination among individuals ≥ aged 65 years. This is mostly due to the lower contribution of age to the models when the analysis was limited to older people. Previous models for older people, using traditional cardiovascular risk factors as predictors, also usually had Harrell's C of below 0.65. Since additional free blood tests are available for the elderly in NBPHSP, there is a promise of improving risk prediction for the elderly by adding
more predictors.

There were significant differences in the calibration performances of the original models across different study regions, including both the three submodels and the combined CVD risk model. This finding suggests that previous universal CVD risk prediction models may encounter the same problems when applied to different regions of China, highlighting the importance of model recalibration. The "ideal recalibration" is not easy to achieve because long-term follow-up is required to obtain the observed 10-year risks. The WHO CVD Risk Chart Working Group provided a flexible recalibration method based on cross-sectional population data (i.e., "practical recalibration" in our study). The present study suggests that the recalibration method proposed by WHO is effective. However, there was still a gap between "practical recalibration" and "ideal recalibration". One possible explanation is that the number of participants in each 5-year age group was relatively small for each study region, leading to errors in estimating the mean risk factor levels and annual incidence of diseases. We assume that if a large and representative population was used to estimate the recalibration parameters, the effect of "practical recalibration" would be closer to that of "ideal recalibration". However, further studies are warranted.

This study population was uniquely large. The loss to follow-up was <1% prior to the global censoring date. More than 100,000 CVD events were documented during a follow-up of 11 years. Unspecified stroke accounted for only 1.4% and 1.2% of first stroke events in women and men. These advantages allowed us to derive risk prediction models separately for three CVD subcategories and obtain robust coefficient estimates. Our models have the potential to be widely used because of readily available predictors from the RHRs of China. The recalibration approach employed in this study enables the flexible updating of our models in different regions in China. We have provided a calculator of practical recalibration parameters (eAppendix 1) and statistical codes required to calculate our models (https://github.com/ckbiobank/yang-2022-ckb-cvd-risk) to support model updating.
Several limitations merit consideration. First, independent samples for external validation of our models are best from local populations in which our models are recalibrated to their CVD incidences and risk factor levels and evaluated the predictive performance. However, to our knowledge, the lack of other regional cohorts of middle-aged and older adults with sufficiently large sample size and including more than 10 years of follow-up in China, limit the options for an effective external validation. With the development of more regional cohorts launched by the China Precision Medicine Initiative and regional electronic RHRs system since the middle of 2010s, further validation and updating of our models are warranted. Second, as a population-based cohort that aimed to follow up for a long time, CKB was not designed to be representative of the general population in China. Despite this, the inclusion of a considerable number of participants from different regions across China and with diverse sociodemographic characteristics enables the study to provide robust regression coefficients and be generalizable to broad populations. When the models are applied to other regions, a recalibration process based on local disease rates and risk factor levels helps improve the applicability of the models. Third, SAH accounts for only a small proportion of all HS cases in the Chinese population, with the corresponding proportion of less than 10% in CKB and precluding us from deriving robust models separately for SAH. We chose to present the HS models which did not distinguish between ICH and SAH in our primary analysis since the coefficients for ICH-specific models hardly ever changed.

Based on the large population-based cohort of Chinese adults, we developed 10-year risk prediction models for IHD, IS, HS, and their combined outcome that may be more widely applicable to RHRs in NBPHSP of China. More importantly, this study took account of the high burden of HS and substantial regional differences in absolute risks of major CVD types within China and derived models that can be adapted for different regions of the Chinese population.
Supplement_1 --http://links.lww.com/WNL/B915

eAppendix 1 -- http://links.lww.com/WNL/B916

WNL-2022-200392_coinvestigator_appendix --http://links.lww.com/WNL/B917

References


11. Hemingway H, Feder GS, Fitzpatrick NK, Denaxas S, Shah AD, Timmis AD. Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAI disease research using...


25. Chinese Society of Cardiology of Chinese Medical Association, Cardiovascular


Figures

Figure 1. Study overview

HS, hemorrhagic stroke; IHD, ischemic heart disease; IS, ischemic stroke
Figure 2. Harrell’s C statistics of the CKB-CVD models

Harrell’s C was calculated using an internal-external cross-validation approach in which each study region was left out of the model fit and used in validation in turn. Harrell’s C shown is the result of pooling Harrell’s C from each external study region.

<table>
<thead>
<tr>
<th>CVD subtypes</th>
<th>Participants</th>
<th>Cases</th>
<th>Harrell C (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ischemic heart disease</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>289,464</td>
<td>27,158</td>
<td>0.732 (0.706-0.758)</td>
</tr>
<tr>
<td>Men</td>
<td>200,132</td>
<td>18,789</td>
<td>0.731 (0.710-0.753)</td>
</tr>
<tr>
<td><strong>Ischemic stroke</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>289,464</td>
<td>24,210</td>
<td>0.759 (0.738-0.779)</td>
</tr>
<tr>
<td>Men</td>
<td>200,132</td>
<td>19,437</td>
<td>0.762 (0.745-0.780)</td>
</tr>
<tr>
<td><strong>Hemorrhagic stroke</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>289,464</td>
<td>5,352</td>
<td>0.803 (0.778-0.827)</td>
</tr>
<tr>
<td>Men</td>
<td>200,132</td>
<td>5,816</td>
<td>0.793 (0.770-0.816)</td>
</tr>
</tbody>
</table>
Figure 3. Model performance of the combined model before and after recalibration

The "original model" refers to the "combined model" before recalibration. (A) Discrimination performance. (B) Calibration performance. \( \chi^2 \) is the Nam-D'Agnostino test chi-square with nine degrees of freedom. The 95% CIs of the observed 10-year risk (black error bar) were too narrow to display in the calibration plots clearly.

<table>
<thead>
<tr>
<th>Models</th>
<th>Participants</th>
<th>Cases</th>
<th>Harrell C (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>289,464</td>
<td>48,845</td>
<td>0.734 (0.732-0.736)</td>
</tr>
<tr>
<td>Men</td>
<td>200,132</td>
<td>37,619</td>
<td>0.743 (0.741-0.746)</td>
</tr>
<tr>
<td><strong>Practical recalibration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>289,464</td>
<td>48,845</td>
<td>0.754 (0.752-0.756)</td>
</tr>
<tr>
<td>Men</td>
<td>200,132</td>
<td>37,619</td>
<td>0.764 (0.762-0.766)</td>
</tr>
<tr>
<td><strong>Ideal recalibration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>289,464</td>
<td>48,845</td>
<td>0.774 (0.772-0.776)</td>
</tr>
<tr>
<td>Men</td>
<td>200,132</td>
<td>37,619</td>
<td>0.771 (0.769-0.774)</td>
</tr>
</tbody>
</table>
Figure 4. Relative IDI (%) of other predictors based on the CKB-CVD models
CVD, cardiovascular disease; MET, metabolic equivalent; rIDI, relative integrated discrimination improvement; Y/N, yes or no. The relative IDI was calculated with the same method used in the predictor selection component (eMethods in the supplement).

a The detailed categories of multi-categorical variables: level of education (6 groups: no formal school, primary school, middle school, high school, technical school or college, and university); smoking status (5 groups: non-smoker, former smoker, current smoker who smoked <10, 10–19, or ≥20 cigarettes or equivalents per day); alcohol consumption (7 groups: non-drinker, former drinker, weekly drinker, daily drinker with an intake of <15, 15–29, 30–59, or ≥60 g/day of pure alcohol); frequency of exercise (4 groups: never or rarely, 1–3 times/month, 1–5 times/week, and daily or almost daily); consumption frequency of fresh fruits, red meat, and eggs (5 groups: daily, 4–6 days/week, 1–3 days/week, monthly, and never or rarely).

b The total physical activity was natural log-transformed before analysis.

c Two participants had missing value of body mass index and were excluded in the current analysis.
Table 1. Summary of available CKB data used in the model derivation

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total participants</td>
<td>289,464</td>
<td>200,132</td>
</tr>
<tr>
<td>Rural area</td>
<td>163,410 (56.5)</td>
<td>114,776 (57.4)</td>
</tr>
<tr>
<td>Age, years</td>
<td>50.4 (42.3–58.1)</td>
<td>51.8 (43.1–60.3)</td>
</tr>
<tr>
<td>Current daily smoker</td>
<td>5874 (2.0)</td>
<td>114,218 (57.1)</td>
</tr>
<tr>
<td>Use of blood pressure-lowering treatment</td>
<td>34,062 (11.8)</td>
<td>20,388 (10.2)</td>
</tr>
<tr>
<td>Systolic blood pressure, mmHg</td>
<td>126.0 (114.0–141.0)</td>
<td>129.5 (119.0–143.0)</td>
</tr>
<tr>
<td>Diastolic blood pressure, mmHg</td>
<td>76.0 (69.0–83.0)</td>
<td>78.5 (71.5–86.0)</td>
</tr>
<tr>
<td>Waist circumference, cm</td>
<td>78.2 (72.0–84.9)</td>
<td>81.3 (74.5–88.5)</td>
</tr>
<tr>
<td>Self-reported diabetes</td>
<td>8258 (2.9)</td>
<td>5055 (2.5)</td>
</tr>
<tr>
<td>Screen-detected diabetes</td>
<td>7933 (2.7)</td>
<td>5169 (2.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cardiovascular outcomes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total person-years observed</td>
<td>3,185,178</td>
<td>2,137,503</td>
</tr>
<tr>
<td>Follow-up time, years</td>
<td>11.2 (10.3–12.2)</td>
<td>11.1 (10.1–12.1)</td>
</tr>
<tr>
<td>Participants with follow-up</td>
<td>241,127 (83.3)</td>
<td>156,488 (78.2)</td>
</tr>
<tr>
<td>≥ 10 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incident events during the first 10 years / total follow-up period</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ischemic heart disease</td>
<td>22,765 / 27,158</td>
<td>15,907 / 18,789</td>
</tr>
<tr>
<td>Hemorrhagic stroke</td>
<td>4482 / 5352</td>
<td>4998 / 5816</td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>19,752 / 24,210</td>
<td>16,398 / 19,437</td>
</tr>
<tr>
<td>Unspecified stroke</td>
<td>1199 / 1522</td>
<td>1112 / 1344</td>
</tr>
<tr>
<td>Total cardiovascular disease</td>
<td>41,152 / 48,845</td>
<td>32,342 / 37,619</td>
</tr>
</tbody>
</table>

Data are n (%) or median (25–75th percentile range) unless otherwise specified.

- Person-year was calculated as the time from the baseline date to the first of the following: death, loss to follow-up, or the global censoring date (December 31, 2017).
- Only the first event was included.
- The number of unspecified stroke events as the first stroke events: 434 in women and 337 in men.
- Total cardiovascular disease events included ischemic heart disease events (fatal and nonfatal I20–I25) and stroke events (fatal or
nonfatal I60–I64).
<table>
<thead>
<tr>
<th></th>
<th><strong>Women</strong></th>
<th></th>
<th><strong>Men</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main effect</td>
<td>Age interaction</td>
<td>Main effect</td>
<td>Age interaction</td>
</tr>
<tr>
<td><strong>Ischemic heart disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age per 5 year</td>
<td>1.40 (1.38–1.41)</td>
<td>–</td>
<td>1.43 (1.41–1.45)</td>
<td>–</td>
</tr>
<tr>
<td>SBP per 10 mm Hg</td>
<td>1.02 (1.01–1.03)</td>
<td>1.00 (0.99–1.00)</td>
<td>1.05 (1.04–1.07)</td>
<td>1.00 (1.00–1.01)</td>
</tr>
<tr>
<td>DBP per 10 mm Hg</td>
<td>1.05 (1.04–1.07)</td>
<td>0.99 (0.99–1.00)</td>
<td>1.10 (1.07–1.12)</td>
<td>0.97 (0.96–0.97)</td>
</tr>
<tr>
<td>Use of blood pressure-lowering treatment</td>
<td>1.55 (1.49–1.61)</td>
<td>0.95 (0.93–0.97)</td>
<td>1.59 (1.52–1.67)</td>
<td>0.97 (0.95–0.99)</td>
</tr>
<tr>
<td>Current daily smoker</td>
<td>1.17 (1.08–1.28)</td>
<td>0.97 (0.93–1.00)</td>
<td>1.22 (1.19–1.26)</td>
<td>0.98 (0.97–1.00)</td>
</tr>
<tr>
<td>Self-reported diabetes</td>
<td>1.59 (1.49–1.69)</td>
<td>0.94 (0.92–0.97)</td>
<td>1.82 (1.69–1.95)</td>
<td>0.91 (0.88–0.94)</td>
</tr>
<tr>
<td>Waist circumference per 10 cm</td>
<td>1.16 (1.14–1.17)</td>
<td>0.97 (0.96–0.98)</td>
<td>1.13 (1.11–1.15)</td>
<td>0.97 (0.96–0.98)</td>
</tr>
<tr>
<td>Baseline survival estimate at 10 years&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.900</td>
<td>–</td>
<td>0.929</td>
<td>–</td>
</tr>
<tr>
<td><strong>Ischemic stroke</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age per 5 year</td>
<td>1.45 (1.43–1.46)</td>
<td>–</td>
<td>1.52 (1.50–1.54)</td>
<td>–</td>
</tr>
<tr>
<td>SBP per 10 mm Hg</td>
<td>1.09 (1.08–1.10)</td>
<td>0.98 (0.97–0.98)</td>
<td>1.13 (1.12–1.15)</td>
<td>0.98 (0.97–0.98)</td>
</tr>
<tr>
<td>DBP per 10 mm Hg</td>
<td>1.11 (1.09–1.13)</td>
<td>0.99 (0.99–1.00)</td>
<td>1.15 (1.13–1.17)</td>
<td>0.98 (0.98–0.99)</td>
</tr>
<tr>
<td>Use of blood pressure-lowering treatment</td>
<td>1.54 (1.48–1.60)</td>
<td>0.96 (0.94–0.97)</td>
<td>1.59 (1.52–1.67)</td>
<td>0.94 (0.93–0.96)</td>
</tr>
<tr>
<td>Current daily smoker</td>
<td>1.30 (1.18–1.42)</td>
<td>0.93 (0.89–0.96)</td>
<td>1.29 (1.24–1.33)</td>
<td>0.97 (0.96–0.99)</td>
</tr>
<tr>
<td>Self-reported diabetes</td>
<td>1.82 (1.71–1.94)</td>
<td>0.90 (0.87–0.93)</td>
<td>1.83 (1.71–1.97)</td>
<td>0.94 (0.91–0.97)</td>
</tr>
<tr>
<td>Waist circumference per 10 cm</td>
<td>1.11 (1.09–1.12)</td>
<td>0.98 (0.97–0.98)</td>
<td>1.08 (1.06–1.10)</td>
<td>0.99 (0.99–1.00)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Baseline survival estimate at 10 years.
<table>
<thead>
<tr>
<th>Predictor variables</th>
<th>Women</th>
<th>Age interaction</th>
<th>Men</th>
<th>Age interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline survival estimate at 10 years</strong></td>
<td>0.910</td>
<td>–</td>
<td>0.928</td>
<td>–</td>
</tr>
<tr>
<td><strong>Hemorrhagic stroke</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age per 5 year</td>
<td>1.51 (1.47–1.54)</td>
<td>–</td>
<td>1.51 (1.47–1.55)</td>
<td>–</td>
</tr>
<tr>
<td>SBP per 10 mm Hg</td>
<td>1.16 (1.14–1.18)</td>
<td>0.98 (0.97–0.99)</td>
<td>1.18 (1.16–1.21)</td>
<td>0.99 (0.98–1.00)</td>
</tr>
<tr>
<td>DBP per 10 mm Hg</td>
<td>1.32 (1.28–1.37)</td>
<td>0.96 (0.95–0.98)</td>
<td>1.33 (1.29–1.38)</td>
<td>0.95 (0.94–0.97)</td>
</tr>
<tr>
<td>Use of blood pressure-lowering treatment</td>
<td>1.62 (1.50–1.74)</td>
<td>0.96 (0.92–0.99)</td>
<td>1.48 (1.37–1.61)</td>
<td>0.95 (0.92–0.99)</td>
</tr>
<tr>
<td>Current daily smoker</td>
<td>1.28 (1.03–1.59)</td>
<td>0.97 (0.89–1.06)</td>
<td>1.10 (1.04–1.17)</td>
<td>1.00 (0.97–1.02)</td>
</tr>
<tr>
<td>Self-reported diabetes</td>
<td>1.50 (1.28–1.75)</td>
<td>0.95 (0.89–1.03)</td>
<td>1.45 (1.22–1.72)</td>
<td>0.96 (0.89–1.04)</td>
</tr>
<tr>
<td>Waist circumference per 10 cm</td>
<td>0.92 (0.89–0.95)</td>
<td>0.99 (0.98–1.01)</td>
<td>0.92 (0.89–0.95)</td>
<td>0.98 (0.97–0.99)</td>
</tr>
<tr>
<td><strong>Baseline survival estimate at 10 years</strong></td>
<td>0.986</td>
<td>–</td>
<td>0.983</td>
<td>–</td>
</tr>
</tbody>
</table>

DBP, diastolic blood pressure; SBP, systolic blood pressure.

Predictor variables were centered as follows: age at 55 years, SBP at 120 mmHg, DBP at 80 mmHg, and waist circumference at 80 cm.

*The Cox models were stratified by ten regions. Baseline survival estimate at 10 years ($S_0(10)$) was estimated by pooling the $S_0(10)$ across regions weighted by the number of events by 10 years.*
Development of Model to Predict 10-Year Risk of Ischemic and Hemorrhagic Stroke and Ischemic Heart Disease Using the China Kadoorie Biobank

Songchun Yang, Yuting Han, Canqing Yu, et al.

Neurology published online April 11, 2022
DOI 10.1212/WNL.0000000000200139

This information is current as of April 11, 2022

Updated Information & Services
including high resolution figures, can be found at:
http://n.neurology.org/content/early/2022/04/11/WNL.0000000000200139.full

Subspecialty Collections
This article, along with others on similar topics, appears in the following collection(s):
Infarction
http://n.neurology.org/cgi/collection/infarction

Permissions & Licensing
Information about reproducing this article in parts (figures, tables) or in its entirety can be found online at:
http://www.neurology.org/about/about_the_journal#permissions

Reprints
Information about ordering reprints can be found online:
http://n.neurology.org/subscribers/advertise