






Associations of polygenic risk scores with risks of stroke and its subtypes in Chinese

Songchun Yang ^{1,2}, Zhijia Sun,¹ Dong Sun,¹ Canqing Yu,^{1,3,4} Yu Guo,⁵ Dianjianyi Sun,^{1,3,4} Yuanjie Pang,^{1,4} Pei Pei ³, Ling Yang,^{6,7} Iona Y Millwood,^{6,7} Robin G Walters,^{6,7} Yiping Chen,^{6,7} Huidong Du,^{6,7} Yan Lu,⁸ Sushila Burgess,⁷ Daniel Avery ⁷, Robert Clarke ⁷, Junshi Chen,⁹ Zhengming Chen,⁷ Liming Li,^{1,3,4} Jun Lv ^{1,3,4,10} On behalf of the China Kadoorie Biobank Collaborative Group

To cite: Yang S, Sun Z, Sun D, et al. Associations of polygenic risk scores with risks of stroke and its subtypes in Chinese. *Stroke & Vascular Neurology* 2024;**9**: e002428. doi:10.1136/svn-2023-002428

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/svn-2023-002428>).

Received 28 February 2023
Accepted 11 August 2023
Published Online First
28 August 2023

ABSTRACT

Background and purpose Previous studies, mostly focusing on the European population, have reported polygenic risk scores (PRSs) might achieve risk stratification of stroke. We aimed to examine the association strengths of PRSs with risks of stroke and its subtypes in the Chinese population.

Methods Participants with genome-wide genotypic data in China Kadoorie Biobank were split into a potential training set (n=22 191) and a population-based testing set (n=72 150). Four previously developed PRSs were included, and new PRSs for stroke and its subtypes were developed. The PRSs showing the strongest association with risks of stroke or its subtypes in the training set were further evaluated in the testing set. Cox proportional hazards regression models were used to estimate the association strengths of different PRSs with risks of stroke and its subtypes (ischaemic stroke (IS), intracerebral haemorrhage (ICH) and subarachnoid haemorrhage (SAH)).

Results In the testing set, during 872 919 person-years of follow-up, 8514 incident stroke events were documented. The PRSs of any stroke (AS) and IS were both positively associated with risks of AS, IS and ICH (p<0.05). The HR for per SD increment (HR_{SD}) of PRS_{AS} was 1.10 (95% CI 1.07 to 1.12), 1.10 (95% CI 1.07 to 1.12) and 1.13 (95% CI 1.07 to 1.20) for AS, IS and ICH, respectively. The corresponding HR_{SD} of PRS_{IS} was 1.08 (95% CI 1.06 to 1.11), 1.08 (95% CI 1.06 to 1.11) and 1.09 (95% CI 1.03 to 1.15). PRS_{ICH} was positively associated with the risk of ICH (HR_{SD}=1.07, 95% CI 1.01 to 1.14). PRS_{SAH} was not associated with risks of stroke and its subtypes. The addition of current PRSs offered little to no improvement in stroke risk prediction and risk stratification.

Conclusions In this Chinese population, the association strengths of current PRSs with risks of stroke and its subtypes were moderate, suggesting a limited value for improving risk prediction over traditional risk factors in the context of current genome-wide association study under-representing the East Asian population.

INTRODUCTION

Stroke is one of the leading causes of death and disease burdens globally.¹ Stroke includes two main subtypes, such as ischaemic stroke

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Polygenic risk scores (PRSs) might achieve risk stratification of stroke.
- ⇒ Evidence from the East Asian population (including Chinese) is lacking.

WHAT THIS STUDY ADDS

- ⇒ The association strengths of current PRSs with risks of stroke and its subtypes were moderate in the Chinese population.
- ⇒ PRS for ischaemic stroke was positively associated with the risk of intracerebral haemorrhage.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ In the Chinese population, current PRSs might have limited value for improving stroke risk prediction over traditional risk factors.
- ⇒ Further studies are warranted to assess whether new PRSs based on larger genome-wide association study or other developing methods have considerable potential to translate into population health benefits.

(IS) and haemorrhagic stroke (HS). The latter could further be divided into intracerebral haemorrhage (ICH) and subarachnoid haemorrhage (SAH). With the accumulation of genomic data worldwide, the genetic background of stroke and its subtypes is gradually being revealed. Polygenic risk score (PRS), a method used to combine minor genetic effects across the whole genome, has been increasingly used in stroke research. Several studies based on European populations have developed PRSs for any stroke (AS) or IS and suggested their potential to improve risk prediction and risk stratification.^{2–9} The incidence of stroke in China, especially ICH, is higher than in Western countries.¹ Recently, a PRS for AS was developed based on the Chinese population and showed similar



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Jun Lv; lvjun@bjmu.edu.cn

association strength in predicting the risk of IS and HS.¹⁰ However, IS and HS might have different aetiological mechanisms.^{11–13} Different stroke subtypes also have their specific genetic loci.¹⁴ No study has specifically developed PRSs for subtypes of stroke in the Chinese population.

The present study was based on a subcohort with genomic data from the China Kadoorie Biobank (CKB). We aimed to examine the association strengths of PRSs with risks of stroke and its subtypes in the Chinese population.

METHODS

Participants

CKB is an ongoing prospective study with 512 724 participants aged 30–79 enrolled from five urban and five rural regions in China between 2004 and 2008. Details of the study have been described elsewhere.¹⁵

Among all CKB participants, there are 100 639 participants with genome-wide genotypic data. Of them, 24 657 participants were selected based on a

case-control design nested within the cohort with the primary aim of studying CVD ('case-control samples'), which formed four matched-case-control training sets (figure 1A, online supplemental methods, tables 1 and 2). The other 75 982 participants were randomly selected from the entire CKB cohort ('population-based samples'); after excluding participants with self-reported coronary artery disease or stroke or transient ischaemic attack at baseline (n=3832), the remaining participants were used as a 'testing set' (n=72 150) (figure 1A, online supplemental methods).

Study design

The current study can be divided into four parts (figure 1B). (1) Validation of previous PRSs. Four previously reported stroke-related PRSs were selected for validation.^{2 4 5 10} (2) Development of new PRSs. Clumping and thresholding ('C+T') and LDpred¹⁶ were used to develop new PRSs for stroke and its subtypes based on two genome-wide association studies with large sample

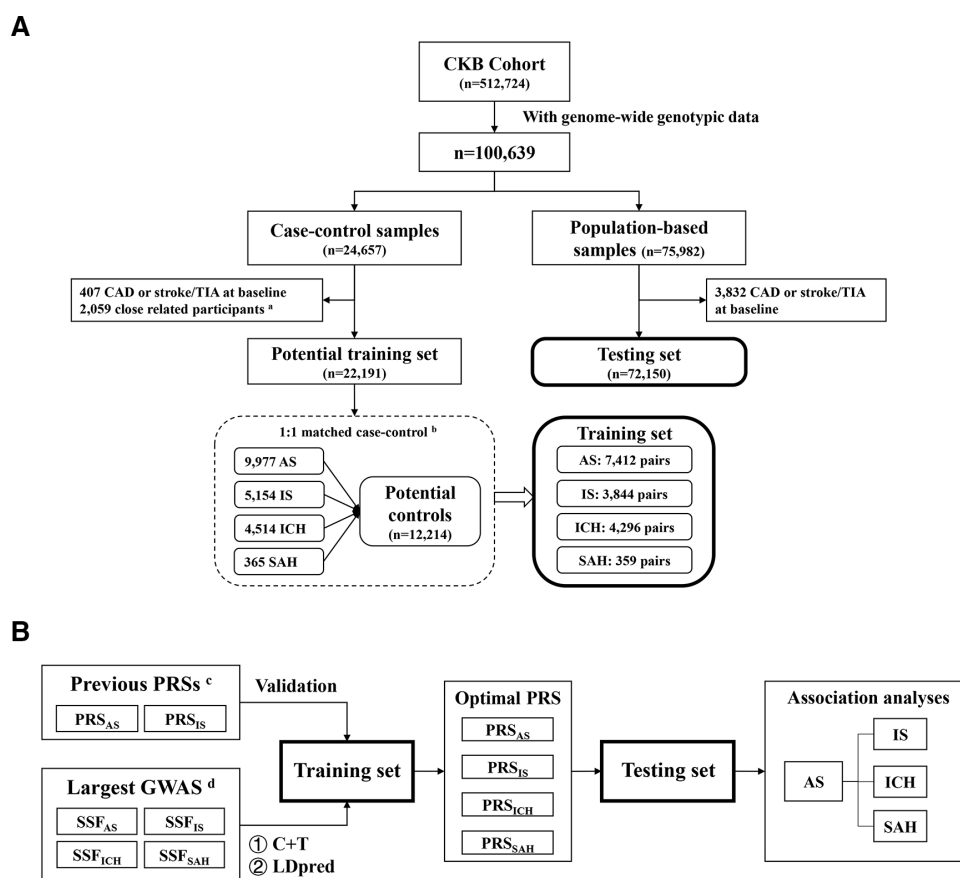


Figure 1 Overview of the present study. (A) Flow chart for the study population; (B) Study design. The current study can be divided into four parts: (1) validation of previous PRSs, (2) development of new PRSs, (3) identification of the optimal PRS for each outcome and (4) validation and evaluation of the optimal PRS for each outcome. ^aParticipants who had a first or second-degree relative in the sample (kinship coefficient $\phi > 0.125$) were removed by using PLINK 1.9. ^bPlease refer to online supplemental methods for detailed procedures of case-control matching. ^cSee online supplemental methods and table 3 for details. ^dSee online supplemental methods and table 4 for details. AS, any stroke; C+T, clumping and thresholding; CAD, coronary heart disease; CKB, China Kadoorie Biobank; GWAS, genome-wide association study; ICH, intracerebral haemorrhage; IS, ischaemic stroke; PRS, polygenic risk score; SAH, subarachnoid haemorrhage; SSF, summary statistics file; TIA, transient ischaemic attack.

sizes.^{14 17} (3) Identification of the optimal PRS for each outcome. The performances of different PRSs in predicting each outcome were compared in the corresponding training sets. (4) Validation and evaluation of the optimal PRS for each outcome. We prospectively examined the associations between optimal PRSs and risks of stroke and its subtypes. We evaluated the impact of PRSs on the risk prediction improvement by adding the optimal PRS to traditional risk prediction models in the testing set.

Assessment of traditional stroke risk factors

The baseline questionnaire collected information on sociodemographic characteristics, lifestyle behaviours, dietary habits, and personal and family medical history.¹⁵ Traditional stroke risk factors considered in the present study included sex, age, systolic and diastolic blood pressure (SBP and DBP), smoking, body mass index (BMI), waist circumference, hypertension, diabetes and family history of stroke. Details on the collection and definition of these variables have been described in our previous work.^{18 19}

Genetic data

At baseline, a 10mL random blood sample was collected from each participant. Genotyping and imputation in this study were centrally conducted, with details provided in our previous study.^{19 20} Briefly, two custom-designed single nucleotide polymorphism (SNP) arrays (Affymetrix Axiom CKB array) were used for genotyping. Imputation was performed based on haplotypes derived from the 1000 Genomes Project Phase 3. There were 9.54million genetic variants with high reliability (online supplemental figure 1).

Polygenic risk scores

We searched the PGS Catalogue,²¹ PubMed and Embase. Four previous stroke PRSs were selected for validation analyses (online supplemental methods and table 3).^{2 4 5 10} Meanwhile, we ran gwasfilter to filter genome-wide association studies (GWAS) from the GWAS Catalogue (<https://www.ebi.ac.uk/gwas/>).^{22 23} Based on ethnicity, sample size and accessibility of the summary statistics file (SSF), we finally included one AS SSF, two SAH SSFs, two ICH SSFs and two IS SSFs from two large-scale GWAS (online supplemental methods and table 4).^{14 17} Similar to our latest research,¹⁹ we developed new PRSs by using two methods: clumping and thresholding ('C+T') and LDpred¹⁶ (online supplemental methods).

Ascertainment of stroke outcomes

All participants were followed up for morbidity and mortality since their baseline enrolment. Incident events were identified by linking with local disease and death registries and the national health insurance database and supplemented by active follow-up.¹⁵ In the testing set, only 653 (0.91%) were lost to follow-up before censoring on 31 December 2018. Trained staff

blinded to baseline information coded all events using the International Classification of Diseases, 10th Revision (ICD-10). Incident stroke events during the follow-up were defined as I60–I64, including SAH (I60), ICH (I61), other nontraumatic intracranial haemorrhage (I62), IS (I63) and unspecified stroke (I64). In the testing set, the events coded as I62 and I64 accounted for only 0.9% (n=76) and 3.5% (n=302) of all incident stroke events.

Since 2014, medical records of incident stroke cases have been retrieved and reviewed by qualified cardiovascular specialists blinded to baseline information. According to a previous study,²⁴ by October 2018, the reporting accuracy was 91.7%, 90.4% and 82.7% for IS, ICH and SAH²⁴; the corresponding diagnostic accuracy was 93.1% (including silent lacunar infarction), 98.2% and 98.1%, respectively.²⁴

Identification of the optimal PRS in the training set

In each training set, we used the conditional logistic regression model to measure the association of each PRS with the risk of the corresponding stroke outcome, stratified by the case-control pair, with the top 10 principal components of ancestry (PCA) and array versions as the covariates. We defined the optimal PRS as the PRS with the highest OR per SD, as our previous study did.¹⁹

Validation and evaluation of the optimal PRS in the testing set

In the testing set, we used the Cox regression model to measure the association of optimal PRSs with risks of stroke and stroke subtypes. The model was stratified by sex and ten study regions, with age as the time scale and adjusting for the top 10 PCA and array versions. We further adjusted for SBP, BMI and family history of stroke in sensitivity analyses. We evaluated the proportional hazards assumptions by examining Schoenfeld residuals. Either non-existent or minimal deviations were observed. In subgroup analyses, the tests for multiplicative interaction were performed using likelihood ratio tests by comparing models with and without cross-product terms between the stratifying variable and PRS.

To evaluate the impact of PRS on risk prediction improvement, we defined the 'CKB-CVD models' as the traditional risk prediction models, as our previous study did.¹⁹ The 'CKB-CVD models' distinguish risks of IS and haemorrhagic stroke and have good discrimination without relying on blood lipids.¹⁸ We added the PRS to traditional models to get a 'PRS-enhanced model'. We assessed the discrimination performance by using Harrell's C.²⁵ We used the net reclassification improvement (NRI) and integrated discrimination improvement to evaluate model reclassification before and after the addition of PRS.²⁶

The study adhered to the PRS Reporting Standards and statement Strengthening the reporting of observational studies in epidemiology for cohort studies simultaneously (online supplemental file 2).^{27 28} Analyses were done with

Table 1 The optimal PRSs associated with risks of stroke and its subtypes in the training sets

Outcomes	Method	PRS source*	No of variants	OR _{SD} (95% CI)	P value	Note
Any stroke (N=7412 pairs)						
	Previous study	PGS002259	448	1.13 (1.09 to 1.16)	1.44×10 ⁻¹¹	
	C+T	GCST005838 (p=1×10 ⁻⁶ , r ² =0)	38	1.11 (1.07 to 1.14)	1.90×10 ⁻⁹	
	LDpred	GCST005838 (p=0.01, Ref=1KGP-EAS)	1 017 531	1.14 (1.10 to 1.18)	3.38×10 ⁻¹⁴	Optimal
Ischaemic stroke (N=3844 pairs)						
	Previous study	PGS000039	1 563 569	1.07 (1.01 to 1.12)	0.012	
	C+T	GCST90018864 (p=0.02, r ² =0.8)	32 158	1.18 (1.13 to 1.24)	3.55×10 ⁻¹¹	Optimal
	LDpred	GCST90018864 (p=0.01, Ref=1KGP-EUR)	1 017 672	1.17 (1.11 to 1.23)	1.46×10 ⁻⁹	
Intracerebral haemorrhage (N=4296 pairs)						
	C+T	GCST90018870 (p=0.001, r ² =0.2)	1326	1.09 (1.04 to 1.14)	1.37×10 ⁻⁴	
	LDpred	GCST90018870 (p=0.1, Ref=1KGP-EUR)	1 017 664	1.10 (1.05 to 1.15)	3.09×10 ⁻⁵	Optimal
Subarachnoid haemorrhage (N=359 pairs)						
	C+T	GCST90018703 (p=0.4, r ² =0)	7899	1.25 (1.06 to 1.47)	9.21×10 ⁻³	Optimal
	LDpred	GCST90018923 (p=0.01, Ref=1KGP-EUR)	1 017 665	1.15 (0.98 to 1.35)	0.096	

The current table only displays the optimal PRS obtained from different strategies (previous study, C+T and LDpred) for each disease outcome. The detailed results of all PRSs can be found in online supplemental table 7.

*'PGS' indicates the index in the PGS Catalogue. 'GCST' indicates the index in the GWAS Catalogue. The information in brackets is the parameter used for developing the PRS.

C+T, clumping and thresholding; EAS, East Asian; EUR, European; 1KGP, 1000 Genomes Project (Phase 3); PRS, polygenic risk score; Ref, reference population.

Stata (V.17.0, StataCorp) and R (V.4.0.3). All statistical tests were two sided with $\alpha=0.05$.

RESULTS

Selection of the optimal PRSs in the training sets

In this study, four 1:1 matched training sets were defined to identify the optimal PRS for AS (7412 pairs), IS (3844 pairs), ICH (4296 pairs) and SAH (359 pairs) (figure 1, online supplemental methods). Among the training sets, 72.7%, 61.6%, 77.9% and 63.8% of the participants were from rural areas in China; 51.9%, 50.5%, 53.4% and 38.4% of the participants were men, respectively. Among the cases, the median age of disease onset (25th–75th percentile) was 65.3 (57.0–72.0), 64.1 (56.1–70.6), 65.9 (57.7–73.0) and 61.0 (53.8–69.2) years, respectively. Among all training sets, the proportion of the control group using the first version of the SNP array was lower than that of the case group (p<0.001) (online supplemental table 2). The performance of PRS for AS and IS developed in previous studies was not better than that of the newly developed PRS in the present study (table 1, online supplemental table 5). The optimal PRS for AS came from the LDpred method, and the optimal PRS for IS, ICH and SAH came from the C+T method. The OR_{SD} (95% CI) of the optimal PRSs was 1.14 (1.10 to 1.18) for AS, 1.18 (1.13 to 1.24) for IS, 1.10 (1.05 to 1.15) for ICH and 1.25 (1.06 to 1.47) for SAH (table 1, online supplemental table 5).

Associations of PRSs with stroke and its subtypes in the testing set

The testing set included 72 150 Chinese participants, of which 59.8% were women. The median age was 50.6 years in women and 51.9 years in men. During 872 919 person-years of follow-up (over 12 years on average), 8514 incident stroke events were documented, including 7507 IS, 1193 ICH and 132 SAH (table 2). The correlations among the optimal PRSs were weak (all correlation coefficients<0.2) (online supplemental figure 2).

The PRS_{AS} and PRS_{IS} were both positively associated with risks of AS, IS and ICH (p<0.05). The HR_{SD} (95% CIs) of PRS_{AS} was 1.10 (1.07 to 1.12), 1.10 (1.07 to 1.12) and 1.13 (1.07 to 1.20) for AS, IS and ICH, respectively. The corresponding HR_{SD} (95% CIs) of PRS_{IS} was 1.08 (1.06 to 1.11), 1.08 (1.06 to 1.11) and 1.09 (1.03 to 1.15) (figure 2, online supplemental table 6). PRS_{ICH} was positively associated with the risk of ICH in the whole testing set (HR_{SD}=1.07), though it was not statistically significant in women (p for sex interaction=0.056) (figure 2C). PRS_{SAH} was not associated with risks of any outcomes (figure 2). A strong association of PRS_{AS} with the risk of SAH (HR_{SD}=1.38, 95% CI 1.03 to 1.87) was observed in men but not in women (p for sex interaction=0.055) (figure 2D).

In sensitivity analyses, the associations of PRSs with risks of stroke and its subtypes did not change significantly after additional adjustment for SBP, BMI and family history of stroke (online supplemental table 6). In subgroup analyses, there was no strong evidence supporting a different association strength across subgroups for IS and ICH after

Table 2 Characteristics of the testing set

	Women	Men
No of participants	43 170	28 980
Baseline characteristics		
Age, years	50.6 (42.5–58.3)	51.9 (43.2–60.3)
Rural areas	22 449 (52.0)	15 772 (54.4)
Array 1	5948 (13.8)	4503 (15.5)
Primary school and below	23 605 (54.7)	11 882 (41.0)
Daily smokers	915 (2.1)	16 317 (56.3)
Body mass index, kg/m ²	23.6 (21.4–26.0)	23.3 (21.1–25.7)
Waist circumference, cm	78.0 (72.0–84.5)	81.5 (74.5–88.5)
Hypertension	14 062 (32.6)	10 653 (36.8)
Diabetes	2 477 (5.7)	1 553 (5.4)
Family history of stroke	7 619 (17.6)	5 075 (17.5)
Follow-up		
Follow-up time, years	12.6 (11.7–13.4)	12.4 (11.4–13.3)
Total person-years*	529 498	343 421
Incident events†		
Any stroke	4 763 (11.0)	3 751 (12.9)
Ischaemic stroke	4 254 (9.9)	3 253 (11.2)
Intracerebral haemorrhage	600 (1.4)	593 (2.0)
Subarachnoid haemorrhage	87 (0.2)	45 (0.2)

Data are presented as n (%) or median (25th–75th percentile) unless otherwise specified.
 *Person-years were calculated as the time from the baseline date to the first of the following: death, lost to follow-up or the global censoring date (31 December 2018).
 †Only the first event was counted.

considering multiple testing (p for interaction $>0.05/8$) (online supplemental figures 3 and 4).

Addition of the optimal PRS to traditional risk prediction models

Based on the traditional models defined in this study, the addition of the PRS did not improve or only slightly

improve the discrimination performance of the models. For IS, the addition of PRS_{AS} increased Harrell's C by 0.0010 in men ($p=0.002$). For haemorrhagic stroke, the addition of PRSs did not influence Harrell's C significantly ($p>0.05$) (figure 3). The addition of the PRS offered little to no improvement in stroke risk stratification. For example, the categorical NRIs at the 10% high-risk threshold for ischaemic and haemorrhagic stroke were all not significant in both sexes ($p>0.05$) (online supplemental table 7).

DISCUSSION

Based on the largest biobank in the Chinese population, only moderate associations were observed between PRSs and risks of stroke and its subtypes in this Chinese population, with an HR_{SD} of about 1.10. The addition of current PRSs offered little to no improvement in stroke risk prediction and risk stratification. We also found that the PRSs developed from GWAS summary statistics of IS were positively associated with the risk of ICH.

In the present study, the associations of PRSs with risks of stroke and its subtypes were moderate, suggesting a limited value for improving risk prediction over traditional risk factors. The HR_{SD} for PRS was usually greater than 1.20 in previous studies of the general population. A PRS for IS (PGS000039) that was developed with the metaGRS method and combined PRSs of 5 stroke subtypes and 14 stroke-related traits had an HR_{SD} of 1.26 (95% CI 1.22 to 1.31) in the European population.⁵ Another PRS for stroke (PGS002259) was also developed using the metaGRS method in a Chinese population, with the HR_{SD} for stroke being 1.28 (95% CI 1.21 to 1.36).¹⁰ However, these two PRSs showed much weaker associations with the risk of stroke or IS in the present study than in previous studies. Since both PRSs were developed using the elastic-net logistic regression, a machine learning approach, the potential overfitting may undermine their generalisation performance.

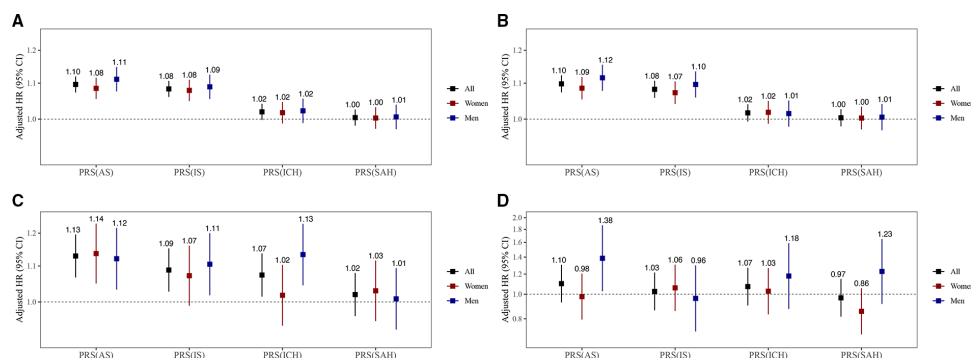


Figure 2 Associations of PRSs with risks of stroke and its subtypes. (A) AS, (B) IS, (C) ICH, (D) SAH. The PRSs reported here are the optimal PRSs for stroke and its subtypes in the training sets (see table 1), which were standardised (0 mean, unit SD) in the testing set. Cox models were stratified by sex and 10 study regions and adjusted for the top 10 principal components of ancestry and array versions, with age as the time scale. The number above the closed square represents the HR. The number of stroke events in women and men has been reported in table 2. The vertical lines indicate 95% CIs. AS, any stroke; ICH, intracerebral hemorrhage; IS, ischaemic stroke; PRS, polygenic risk score; SAH, subarachnoid haemorrhage.

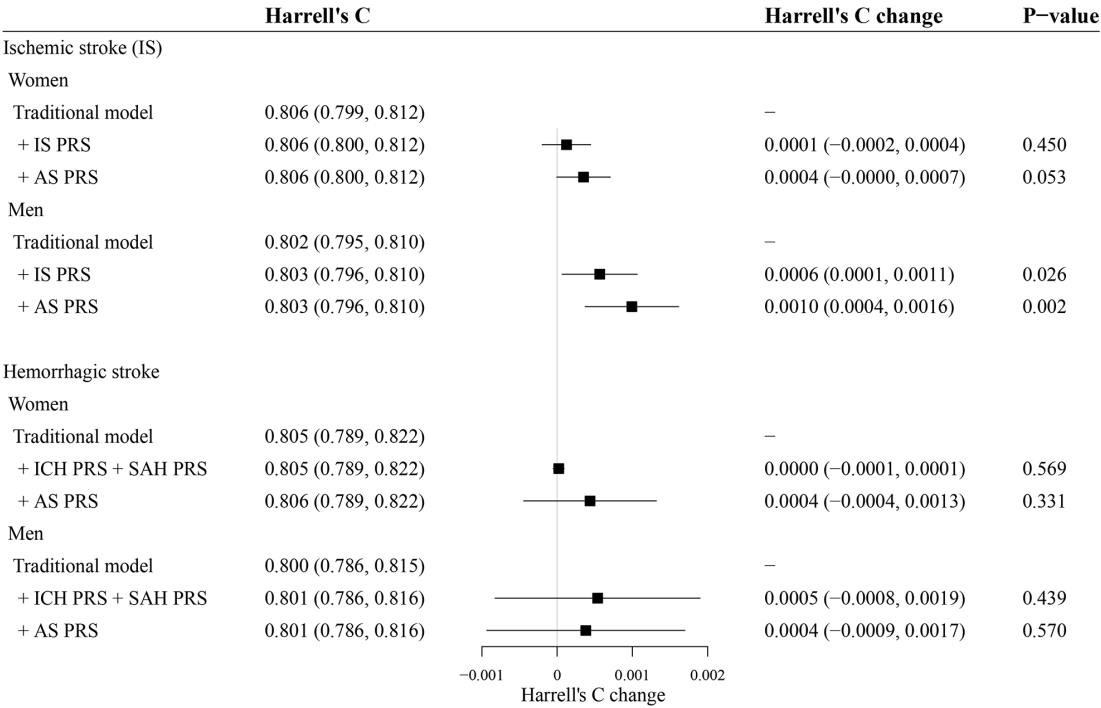


Figure 3 C statistics evaluating the performance of PRS. The traditional risk prediction models (traditional models) were defined as sex-specific Cox models stratified by 10 study regions, with time on study as the time scale, including models for ischaemic stroke (ICD-10: I63) and models for haemorrhagic stroke (ICD-10: I60–I62).¹⁸ Predictors included in traditional models were the same as the ‘CKB-CVD models’, including age, systolic and diastolic blood pressure, use of antihypertensives, current daily smoking, self-reported diabetes and waist circumference. Interactions between age and the other six predictors were also included. The 95% CIs of Harrell’s C and Harrell’s C changes were calculated by 100 bootstrap replications using the BCa method in Stata. CKB, China Kadoorie Biobank; CVD, Cardiovascular disease; ICD, International Classification of Disease; PRS, Polygenic risk score.

The incidence rate of ICH is much higher in Chinese than in European populations. However, non-European populations are under-represented in GWAS, which serves as the basis for PRS development. The largest GWAS for ICH included only 3400 ICH cases, with most of them from European populations.¹⁷ The present study attempted to develop PRS for ICH based on summary statistics from this GWAS. The weak associations observed in the present study are either explained by the difference in genetic background between ethnic groups or suggest that this GWAS may be underpowered. The stronger association estimate between PRS and HS risk reported in the previous study was likely due to the inclusion of PRSs for risk factors of HS (such as blood pressure) in the metaGRS method.¹⁰ It is worth mentioning that, in the present study, the PRSs directly developed from GWAS summary statistics of IS were also positively associated with the risk of ICH. Although there are differences in aetiology and risk factor profile between IS and ICH,^{11–13} they might also have some partially shared aetiological mechanisms like the cerebral small-vessel disease.²⁹

This study has the following strengths. The large sample size and a large number of stroke events (including IS and ICH) enabled us to separate powerful training sets and the testing set and to conduct subgroup analyses. The lost to follow-up rate was less than 1% at an average follow-up period of over 12 years in CKB. The main subtypes of

stroke (ie, IS, ICH and SAH) were well classified, and the reporting and diagnostic accuracy of stroke events were high.²⁴ The genotyping and imputation of genetic data in this study were centrally conducted through a standard quality control process. Genetic variants with high reliability covered the whole genome well.

However, several limitations merit consideration. First, we did not further consider the subtypes of IS (eg, large-atherosclerotic stroke, cardioembolic stroke and small vessel stroke) as over 75% of the incident IS events were coded as unspecified IS (ICD-10: I63.9), which precluded us from conducting more detailed analyses. Previous studies have suggested that there are differences in genetic loci of different IS subtypes.^{14 30} Subsequent studies can explore whether distinguishing IS subtypes can further improve the predictive ability of PRS for IS. Second, compared with IS and ICH, the number of SAH events was relatively small. Therefore, it is difficult to exclude chance factors for the positive results observed in the present study. Further studies with more SAH events are warranted to examine our findings. Third, the genetic variants with ambiguous SNP (ie, A/T, C/G) and those that were not found in CKB or had low imputation quality scores were removed during the standard quality control process of PRSs. This might weaken the associations of previous PRSs with stroke and its subtypes. Fourth, because information on blood lipids was not available for

the current study population, we were unable to compare the impacts of blood lipids and PRS on traditional stroke risk prediction model improvement. However, the addition of blood lipids may enhance the traditional non-laboratory-based models, as previous studies have shown.^{31 32} Therefore, adding PRS to a 'lipid-enhanced model' might lead to a more minor improvement than what we have observed in the present study.

CONCLUSIONS

In this Chinese population, the associations of optimal PRSs with risks of stroke and its subtypes were moderate, suggesting a limited value for improving risk prediction over traditional risk factors in the context of current GWAS under-representing the East Asian population. As GWAS of stroke and its subtypes progress among East Asians, further studies are warranted to assess whether new PRSs have considerable potential to translate into precision public health and population health benefits and, if so, to determine the appropriate context for their use.

Author affiliations

¹Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing, China

²Department of Dermatology, Xiangya Hospital, Central South University, Changsha, Hunan, China

³Peking University Center for Public Health and Epidemic Preparedness & Response, Beijing, China

⁴Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China

⁵Fuwai Hospital Chinese Academy of Medical Sciences, Beijing, China

⁶Medical Research Council Population Health Research Unit at the University of Oxford, Oxford, UK

⁷Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford, UK

⁸NCDs Prevention and Control Department, Suzhou CDC, Suzhou, Jiangsu, China

⁹China National Center for Food Safety Risk Assessment, Beijing, China

¹⁰State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing, China

Acknowledgements The most important acknowledgment is to the participants in the study and the members of the survey teams in each of the 10 regional centres, as well as to the project development and management teams based in Beijing, Oxford and the 10 regional centres.

Collaborators International Steering Committee: Junshi Chen, Zhengming Chen (PI), Robert Clarke, Rory Collins, Yu Guo, Liming Li (PI), Jun Lv, Richard Peto, Robin Walters. International Co-ordinating Centre, Oxford: Daniel Avery, Derrick Bennett, Ruth Boxall, Sue Burgess, Ka Hung Chan, Yumei Chang, Yiping Chen, Zhengming Chen, Johnathan Clarke, Robert Clarke, Huaidong Du, Ahmed Edris Mohamed, Zamy 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, Sam Morris, 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, Jun Lv, Pei Pei, Chao Liu, Canqing Yu, Qingmei Xia. 10 Regional Co-ordinating Centres: Qingdao CDC: Zengchang Pang, Ruqin Gao, Shansheng Li, Haiping Duan, Shaojie Wang, Yongmei Liu, Ranran Du, Yajing Zang, Liang Cheng, Xiaocao Tian, Hua Zhang, Yaoming Zhai, Feng Ning, Xiaohui Sun, Feifei Li. Licang CDC: Silu Lv, Junzheng Wang, Wei Hou. Heilongjiang Provincial CDC: Wei Sun, Shichun Yan, Xiaoming Cui. Nangang CDC: Chi Wang, Zhenyuan Wu, Yanjie Li, Quan Kang. Hainan Provincial CDC: Huiming Luo, Tingting Ou. Meilan CDC: Xiangyang Zheng, Zhendong Guo, Shukuan Wu, Yilei Li, Huimei Li. Jiangsu Provincial CDC: Ming Wu, Yonglin Zhou, Jinyi Zhou, Ran Tao, Jie Yang, Jian Su. Suzhou CDC: Fang

Liu, Jun Zhang, Yihe Hu, Yan Lu, Liangcai Ma, Aiyu Tang, Shuo Zhang, Jianrong Jin, Jingchao Liu. Guangxi Provincial CDC: Mei Lin, Zhenzhen Lu. Liuzhou CDC: Lifang Zhou, Changping Xie, Jian Lan, Tingping Zhu, Yun Liu, Liuping Wei, Liyan Zhou, Ningyu Chen, Yulu Qin, Sisi Wang. Sichuan Provincial CDC: Xianping Wu, Ningmei Zhang, Xiaofang Chen, Xiaoyu Chang. Pengzhou CDC: Mingqiang Yuan, Xia Wu, Xiaofang Chen, Wei Jiang, Jiaqiu Liu, Qiang Sun. Gansu Provincial CDC: Faqing Chen, Xiaolan Ren, Caixia Dong. Maiji CDC: Hui Zhang, Enke Mao, Xiaoping Wang, Tao Wang, Xi zhang. Henan Provincial CDC: Kai Kang, Shixian Feng, Huizi Tian, Lei Fan. Huixian CDC: Xiaolin Li, Huarong Sun, Pan He, Xukui Zhang. Zhejiang Provincial CDC: Min Yu, Ruying Hu, Hao Wang. Tongxiang CDC: Xiaoyi Zhang, Yuan Cao, Kaixu Xie, Lingli Chen, Dun Shen. Hunan Provincial CDC: Xiaojun Li, Donghui Jin, Li Yin, Huilin Liu, Zhongxi Fu. Liuyang CDC: Xin Xu, Hao Zhang, Jianwei Chen, Yuan Peng, Libo Zhang, Chan Qu.

Contributors JL conceived and designed the study. LL, ZC and JC: members of the China Kadoorie Biobank Steering Committee, designed and supervised the whole study, obtained funding, and, together with CY, YG, DiS, YP, PP, LY, YC, HD, YL, SB, DA, IYM and RGW: acquired the data. SY, ZS and DoS analysed the data. SY drafted the manuscript. CY, YP, DiS and RC helped to interpret the results. JL contributed to the critical revision of the manuscript for important intellectual content. All authors reviewed and approved the final manuscript. JL is the guarantor.

Funding This work was supported by the 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, 91846303, 81941018) and Chinese Ministry of Science and Technology (2011BAI09B01).

Disclaimer The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and CKB had ethical approvals from the Ethical Review Committee of the Chinese Center for Disease Control and Prevention (Beijing, China) (approval notice: 005/2004) and the Oxford Tropical Research Ethics Committee, University of Oxford (UK) (reference: 025-04). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Details of how to access China Kadoorie Biobank data and details of the data release schedule are available from www.ckbiobank.org/site/Data+Access.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Songchun Yang <http://orcid.org/0000-0002-0021-4178>

Pei Pei <http://orcid.org/0000-0002-5741-6563>

Daniel Avery <http://orcid.org/0000-0002-9823-9575>

Robert Clarke <http://orcid.org/0000-0002-9802-8241>

Jun Lv <http://orcid.org/0000-0001-7916-3870>

REFERENCES

- Feigin VL, Stark BA, Johnson CO, *et al*. Stroke collaborators. global, regional, and national burden of stroke and its risk factors, 1990-

- 2019: a systematic analysis for the global burden of disease study 2019. *Lancet Neurol* 2021;20:795–820.
- 2 Ibrahim-Verbaas CA, Fornage M, Bis JC, *et al*. Predicting stroke through genetic risk functions: the CHARGE risk score project. *Stroke* 2014;45:403–12.
- 3 Malik R, Bevan S, Nalls MA, *et al*. Multilocus genetic risk score associates with ischemic stroke in case-control and prospective cohort studies. *Stroke* 2014;45:394–402.
- 4 Rutten-Jacobs LC, Larsson SC, Malik R, *et al*. Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306473 UK biobank participants. *BMJ* 2018;363:k4168.
- 5 Abraham G, Malik R, Yonova-Doing E, *et al*. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun* 2019;10:5819.
- 6 Li J, Chaudhary DP, Khan A, *et al*. Polygenic risk scores augment stroke subtyping. *Neurol Genet* 2021;7:e560.
- 7 Marston NA, Patel PN, Kamanu FK, *et al*. Clinical application of a novel genetic risk score for ischemic stroke in patients with cardiometabolic disease. *Circulation* 2021;143:470–8.
- 8 O'Sullivan JW, Shcherbina A, Justesen JM, *et al*. Combining clinical and polygenic risk improves stroke prediction among individuals with atrial fibrillation. *Circ Genom Precis Med* 2021;14:e003168.
- 9 Sun L, Pennells L, Kaptoge S, *et al*. Polygenic risk scores in cardiovascular risk prediction: a cohort study and modelling analyses. *PLoS Med* 2021;18:e1003498.
- 10 Lu X, Niu X, Shen C, *et al*. Development and validation of a polygenic risk score for stroke in the Chinese population. *Neurology* 2021;97:e619–28.
- 11 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.
- 12 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.
- 13 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.
- 14 Malik R, Chauhan G, Traylor M, *et al*. Multi-ancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* 2018;50:524–37.
- 15 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.
- 16 Vilhjálmsdóttir BJ, Yang J, Finucane HK, *et al*. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet* 2015;97:576–92.
- 17 Sakaue S, Kanai M, Tanigawa Y, *et al*. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* 2021;53:1415–24.
- 18 Yang S, Han Y, Yu C, *et al*. Development of a model to predict 10-year risk of ischemic and hemorrhagic stroke and ischemic heart disease using the China Kadoorie biobank. *Neurology* 2022;98:e2307–17.
- 19 Yang S, Sun D, Sun Z, *et al*. Minimal improvement in coronary artery disease risk prediction in Chinese population using polygenic risk scores: evidence from the China Kadoorie biobank. *Chin Med J (Engl)* 2023.
- 20 Zhu Z, Li J, Si J, *et al*. A large-scale genome-wide association analysis of lung function in the Chinese population identifies novel loci and highlights shared genetic etiology with obesity. *Eur Respir J* 2021;58:2100199.
- 21 Lambert SA, Gil L, Jupp S, *et al*. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* 2021;53:420–5.
- 22 Yang S, Li C, Hu Y, *et al*. Gwasfilter: an R script to filter genome-wide Association study. *Chin J Epidemiol* 2021;42:1876–81.
- 23 Buniello A, MacArthur JAL, Cerezo M, *et al*. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;47:D1005–12.
- 24 Turnbull I, Clarke R, Wright N, *et al*. Diagnostic accuracy of major stroke types in Chinese adults: a clinical adjudication study involving 40,000 stroke cases. *Lancet Reg Health West Pac* 2022;21:100415.
- 25 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–87.
- 26 Pencina MJ, D'Agostino RB, D'Agostino RB, *et al*. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.
- 27 Wand H, Lambert SA, Tamburro C, *et al*. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 2021;591:211–9.
- 28 Elm E von, Altman DG, Egger M, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806–8.
- 29 Wardlaw JM, Smith C, Dichgans M. Small vessel disease: mechanisms and clinical implications. *Lancet Neurol* 2019;18:684–96.
- 30 Linden AB, Clarke R, Hammami I, *et al*. Genetic associations of adult height with risk of cardioembolic and other subtypes of ischemic stroke: a mendelian randomization study in multiple ancestries. *PLoS Med* 2022;19:e1003967.
- 31 Ueda P, Woodward M, Lu Y, *et al*. Laboratory-based and office-based risk scores and charts to predict 10-year risk of cardiovascular disease in 182 countries: a pooled analysis of prospective cohorts and health surveys. *Lancet Diabetes Endocrinol* 2017;5:196–213.
- 32 Kaptoge S, Pennells L, De Bacquer D, *et al*. The WHO CVD risk chart working group. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *Lancet Glob Health* 2019;7:e1332–45.