

Stroke Risk Evaluation in a Contemporary Population of 0.5 Million Chinese Adults

Matthew Chun

Christ Church

University of Oxford

*This thesis is submitted to the Department of Engineering Science,
University of Oxford, in fulfilment of the requirements for the degree
of
Doctor of Philosophy*

Trinity Term, 2021

Abstract

Stroke is a leading cause of death and disability globally, with China bearing the largest stroke burden in the world. Global strokes can be halved through primary prevention strategies including the use of risk prediction models to identify individuals with high risk of stroke. However, the models advocated for in current international guidelines have been derived mainly in high-income countries, and their clinical utility in a contemporary Chinese population is uncertain. Moreover, these models generally do not account for the heterogeneity between ischaemic stroke (IS) and haemorrhagic stroke (HS) pathological types as well as further aetiological subtypes of IS. The investigations in this thesis describe the assessment of existing risk prediction methods and development of novel models for evaluating primary stroke risk in the China Kadoorie Biobank – a prospective cohort study of 0.5 million Chinese adults recruited from 10 areas in China in 2004-2008. First, we evaluated the 2017 Framingham Stroke Risk Profile in CKB, and used a similar Cox modelling approach to develop novel models for 9-year risk prediction of total stroke that yielded a 6%-8% increase in area under the receiver operating characteristic curve (AUC) and >90% reduction in the calibration χ^2 test statistic. Risk prediction models for IS and HS were developed with similar results. Second, machine learning approaches including logistic regression, support vector machines, random survival forests, gradient boosted trees (GBT), and multilayer perceptrons were compared for 9-year risk prediction of total stroke. We also developed an ensemble approach combining Cox and GBT predictions, which identified individuals at high-risk of stroke with higher accuracy, specificity, and precision than any single-model approach. Third, we explored the utility of single versus sequential measurements of risk factors for prediction of stroke, and did not find any substantial improvements from sequential risk factor measurements. Fourth, we developed models to assess differential risk of IS and HS and estimated reductions in events and fatalities using optimised statin therapy policies informed by such models. Finally, we characterised the heterogeneity of IS aetiological subtypes and developed models to classify 24% of CKB's primary IS events that were previously characterised as having undetermined aetiology by the existing Causative Classification System (CCS) for Ischemic Stroke. The contributions of this thesis advance the understanding and prevention of stroke in China.

Stroke Risk Evaluation in a Contemporary Population of 0.5 Million Chinese Adults



Matthew Chun
Christ Church
University of Oxford

Supervised by
Prof. David Clifton
Dr. Tingting Zhu

This thesis is submitted to the Department of Engineering Science,
University of Oxford, in fulfilment of the requirements for the degree
of

Doctor of Philosophy

Trinity Term, 2021

This thesis is dedicated to
Roy, Margaret, Nicole, and Ares.
Thank you for your love and support.

Declaration

I declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Oxford or any other University or similar institution.

Matthew Chun
February 23, 2022

Acknowledgements

First and foremost, I would like to thank God, who has been the source of every opportunity, talent, and inspiration in my life. This thesis was also made possible by the support of many colleagues, friends, and family members.

I would like to thank Professor David Clifton and Dr. Tingting Zhu for their years of supervision and guidance. I would also like to thank Professor Robert Clarke and Dr. Benjamin Cairns from the Nuffield Department of Population Health for their close collaboration and crucial involvement in my development not only as an engineer, but as an epidemiologist. The best interdisciplinary research lies at the intersection of diverse team players with deep domain expertise, and I consider myself extremely blessed to have spent my years at Oxford in the midst of such a junction.

In addition, I would like to thank the rest of my colleagues in the Computational Health Informatics Laboratory and the China Kadoorie Biobank group. The cross-pollination of ideas, critical evaluation of my research, and many morale-boosting lunches shared at the Big Data Institute have been transformative for both me and my work.

Thank you to the Rhodes Trust for funding this research and for welcoming me into a community of exceptionally talented and caring scholars.

Thank you also to my housemates, neighbours, friends, and church community for their endless support and kindness. I could not have survived the dreary English winters and COVID-19 pandemic without them.

I am further grateful to my family for a lifetime of emotional, intellectual, spiritual, and financial support, which continued unfailingly even from across the Atlantic. I am especially thankful to them for taking me back into their care for several months of the pandemic when I needed them the most.

Last, but not least, I would like to thank my best friend, Ares. This thesis began before his birth and was completed after his death. Consequently, it was the only project we got to collaborate on in his short, but vibrant life. There is a remarkable common latent space between training machine learning models and training golden retrievers, and much credit is due to Ares for enabling this knowledge transfer. Many thanks indeed to Ares for getting me out of bed each morning as well as taste-testing (and tearing to shreds) many manuscript drafts. I am confident that this final product would have received his paw print of approval.

Abstract

Stroke is a leading cause of death and disability globally, with China bearing the largest stroke burden in the world. Global strokes can be halved through primary prevention strategies including the use of risk prediction models to identify individuals with high risk of stroke. However, the models advocated for in current international guidelines have been derived mainly in high-income countries, and their clinical utility in a contemporary Chinese population is uncertain. Moreover, these models generally do not account for the heterogeneity between ischaemic stroke (IS) and haemorrhagic stroke (HS) pathological types as well as further aetiological subtypes of IS. The investigations in this thesis describe the assessment of existing risk prediction methods and development of novel models for evaluating primary stroke risk in the China Kadoorie Biobank – a prospective cohort study of 0.5 million Chinese adults recruited from 10 areas in China in 2004-2008. First, we evaluated the 2017 Framingham Stroke Risk Profile in CKB, and used a similar Cox modelling approach to develop novel models for 9-year risk prediction of total stroke that yielded a 6%-8% increase in area under the receiver operating characteristic curve (AUC) and >90% reduction in the calibration χ^2 test statistic. Risk prediction models for IS and HS were developed with similar results. Second, machine learning approaches including logistic regression, support vector machines, random survival forests, gradient boosted trees (GBT), and multilayer perceptrons were compared for 9-year risk prediction of total stroke. We also developed an ensemble approach combining Cox and GBT predictions, which identified individuals at high-risk of stroke with higher accuracy, specificity, and precision than any single-model approach. Third, we explored the utility of single versus sequential measurements of risk factors for prediction of stroke, and did not find any substantial improvements from sequential risk factor measurements. Fourth, we developed models to assess differential risk of IS and HS and estimated reductions in events and fatalities using optimised statin therapy policies informed by such models. Finally, we characterised the heterogeneity of IS aetiological subtypes and developed models to classify 24% of CKB's primary IS events that were previously characterised as having undetermined aetiology by the existing Causative Classification System (CCS) for Ischemic Stroke. The contributions of this thesis advance the understanding and prevention of stroke in China.

Publications

Thesis-Related Publications

- [1] **Chun M**, Clarke R, Cairns BJ, Clifton D, Zhu T. A Review of Risk Prediction Models for Primary Prevention of Stroke. [**In preparation**]
- [2] **Chun M**, Clarke R, Zhu T, Clifton D, Bennett D, Chen Y, et al. Risk scores for prediction of total stroke and stroke types in Chinese adults: a prospective study of 0.5 million adults. [**Under review**]
- [3] **Chun M**, Clarke R, Cairns BJ, Clifton D, Bennett D, Chen Y, et al. Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *J Am Med Inform Assoc.* 2021.
- [4] **Chun M**, Clarke R, Cairns BJ, Clifton D, Bennett D, Chen Y, et al. Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *8th International Conference on Neurology and Epidemiology.* 2022. (Invited Speaker) [Abstracts to be published in *Neuroepidemiology*]
- [5] **Chun M**, Clarke R, Zhu T, Clifton D, Bennett D, Chen Y, Guo Y, et al. Utility of single versus sequential measurements of risk factors for prediction of stroke in Chinese adults. *Sci Rep.* 2021.
- [6] **Chun M**, Clarke R, Zhu T, Clifton D, Bennett D, Chen Y, et al. Optimal use of statin therapy for primary prevention of stroke and major coronary events: 9-year follow-up of 0.5M adults in the China Kadoorie Biobank. [**In preparation**]
- [7] **Chun M**, Qin H, Turnbull I, Sansome S, Gilbert S, Wright N, et al. Ischemic stroke subtypes in 0.5 million Chinese adults: incidence, outcomes, and classification of cases with undetermined etiology. [**In preparation**]

Other Publications

- [1] Youssef A, ... **Chun M**, et al. Transfer Learning methodologies for COVID-19 and future pandemic preparedness. [**In preparation**]

Contents

List of Figures	xiii
List of Tables	xvii
List of Abbreviations	xx
1 Introduction	1
1.1 Introduction to Stroke in China	1
1.2 Thesis Objectives and Overview	3
2 Background	5
2.1 Pathology of Stroke	5
2.1.1 Stroke Definitions	5
2.1.2 Stroke Pathological Types	6
2.1.3 Subtypes of Ischaemic Stroke	6
2.1.4 Subtypes of Haemorrhagic Stroke	8
2.1.5 ICD-10 Codes for Stroke Pathologies	9
2.2 Primary Prevention of Stroke	10
2.2.1 Risk Factors	10
2.2.2 Population-Based Strategies	12
2.2.3 Risk-Based Strategies	13
2.3 Stroke Risk Models	14
2.3.1 Traditional Approaches for Stroke Risk Prediction	18
2.3.2 Machine Learning Approaches for Stroke Risk Prediction	21
2.4 Conclusion	24
3 The China Kadoorie Biobank	26
3.1 Study Design	26
3.2 Validation and Adjudication of Stroke Endpoints	28
3.3 Selection of CKB Data	29
3.4 Descriptive Statistics of the China Kadoorie Biobank	32
3.4.1 Distributions of Well-Established Stroke Risk Factors	32
3.4.2 Missing Data and Censoring	36
3.5 Conclusion	37

4	Traditional Prediction of 9-Year Stroke Risk	39
4.1	Introduction	39
4.1.1	Survival Analysis and Cox Regression	40
4.2	Methods	42
4.2.1	Inclusion Criteria and Data Preprocessing	42
4.2.2	Model Development	42
4.2.3	Statistical Analysis	44
4.3	Results	45
4.3.1	Baseline Characteristics	45
4.3.2	Assessment and Update of 2017 Framingham Stroke Risk Profile	46
4.3.3	Prediction of Total Stroke After Adjusting for Areas in China	47
4.3.4	Expanded Risk Equations with Additional Risk Factors . . .	48
4.3.5	Risk Equations for Different Stroke Pathological Types . . .	49
4.3.6	Sensitivity Analyses	50
4.4	Discussion	52
4.5	Conclusion	54
5	Machine Learning for Stroke Risk Prediction	56
5.1	Introduction	56
5.1.1	Overview of ML Approaches	57
5.2	Methods	63
5.2.1	Inclusion Criteria	63
5.2.2	Model Development and Validation	64
5.2.3	Statistical Analysis	66
5.3	Results	67
5.3.1	Baseline Characteristics	67
5.3.2	Comparisons of Cox versus ML models to predict risk of stroke	68
5.3.3	Prediction of Best Model for Risk Screening	71
5.3.4	Evaluation of Ensemble Approach for Risk Screening	72
5.4	Discussion	73
5.5	Conclusion	77
6	Single v. Sequential Risk Factor Measurements for Stroke Risk Prediction	78
6.1	Introduction	78
6.2	Methods	79
6.2.1	Inclusion Criteria and Data Preprocessing	79
6.2.2	Model Development	82
6.2.3	Statistical Analysis	85

6.3	Results	86
6.4	Discussion	90
6.5	Conclusion	96
7	Optimal Statin Therapy Strategies for Primary Stroke Prevention	97
7.1	Introduction	97
7.2	Methods	99
7.2.1	Study Design and Participants	99
7.2.2	Statistical Analysis	99
7.3	Results	102
7.3.1	Baseline Characteristics	102
7.3.2	Incidence of Primary Ischaemic Stroke and Haemorrhagic Stroke	102
7.3.3	Discrimination of Ischaemic Versus Haemorrhagic First Stroke Cases	105
7.3.4	Optimal Use of Statin Therapy to Prevent Stroke Events and Stroke Deaths	109
7.3.5	Additional Inclusion of Major Coronary Events	109
7.4	Discussion	112
7.5	Conclusion	115
8	Heterogeneity and Classification of Ischaemic Stroke Subtypes	116
8.1	Introduction	116
8.2	Methods	117
8.2.1	Study Design	117
8.2.2	Data Used for CCS Classification of IS Subtypes	118
8.2.3	Statistical Analysis	120
8.3	Results	122
8.3.1	Heterogeneity of Ischaemic Stroke Subtypes	122
8.3.2	Classification of Ischaemic Stroke Cases	125
8.4	Discussion	128
8.5	Conclusion	130
9	Conclusions and Future Work	131
9.1	Summary of Thesis	132
9.1.1	Key Contributions	136
9.2	Future Work	137
9.2.1	External Validation and Updating of Models	137
9.2.2	Validation of ML-Classified Ischaemic Stroke Subtype Endpoints	137
9.2.3	Translation to Clinical Practice	138
9.3	Closing Thoughts	139

Appendices

A	CKB Risk Factor Indicators and Adjudication Variables	142
B	Chapter 4 Supplementary Materials	147
B.1	Testing of Proportional Hazards Assumption	147
B.2	Additional Figures and Tables	149
C	Chapter 5 Supplementary Materials	155
C.1	Details of Cox and ML Model Development	155
C.1.1	Cox Models	155
C.1.2	Random Survival Forest (RSF) Models	155
C.1.3	Logistic Regression (LR) Models	156
C.1.4	Support Vector Machine (SVM) Models	156
C.1.5	Gradient Boosted Tree (GBT) Models	157
C.1.6	Multilayer Perceptron (MLP) Models	158
C.1.7	Decision Tree for Selection between Cox and GBT	161
C.2	Additional Figures and Tables	162
D	Chapter 6 Supplementary Materials	166
D.1	Hyperparameter Tuning of GBT Models	166
D.1.1	GBT: Single measurement at most recent visit	167
D.1.2	GBT: Sequential measurements at three visits	167
D.1.3	GBT: Longitudinal summary of sequential measurements at three visits	167
D.1.4	GBT: Longitudinal summary of stroke risk estimates at three visits + Single measurement at most recent visit	168
D.2	Additional Figures and Tables	168
E	Chapter 7 Supplementary Materials	175
E.1	Details of GBT Model Development	175
E.2	Additional Figures and Tables	176
F	Chapter 8 Supplementary Materials	180
F.1	Implementation of Causative Classification System for Ischemic Stroke in CKB	180
F.1.1	Fitting of CKB Variables to CCS	180
F.1.2	Limitations of Using CCS in CKB	182
F.2	Details of Multiclass LR Model Development	182
F.2.1	Data Release	182
F.2.2	Data Preprocessing	182
F.2.3	Multiclass LR Model Construction	183
F.3	Additional Figures and Tables	184

Contents

xii

References

188

List of Figures

2.1	Pathological types and subtypes of stroke.	7
3.1	Map of CKB regions.	27
3.2	Breakdown of stroke pathologies in CKB.	30
3.3	Age distributions by sex.	32
3.4	Systolic blood pressure distributions by sex and by use of blood pressure medication.	33
3.5	Diabetes mellitus prevalence by sex.	34
3.6	Prior CHD prevalence by sex.	34
3.7	Distributions of current smoking behaviour by sex.	35
3.8	Urban vs rural demographics by sex.	36
3.9	Censoring in CKB.	37
4.1	Flowchart of Cox model development process.	44
4.2	Calibration of the 2017 Framingham Stroke Risk Profile, and recalibrated and refitted models from CKB, for total stroke in men and women.	46
4.3	Multivariable hazard ratios (HRs) and 95% CI for total stroke in men and women, for Framingham and for CKB.	48
4.4	Area-specific baseline survival curves and 9 year incidence for IS and HS in CKB.	49
4.5	Predicted risk distributions for IS and HS by CKB area.	50
4.6	Comparison of 9-year risk predictions in test set for Cox model (“+Refitting”) and Fine-Gray model for total stroke, IS, and HS in men and women.	51
5.1	Cox and ML model development and validation.	64
5.2	Distribution of strokes over time from baseline survey in 503,842 adults with no prior history of stroke in CKB.	67
5.3	Calibration plots for the Cox and ML models in men and women.	69
5.4	Discrimination and calibration for the Cox and ML models for risk prediction of stroke at various time scales.	70

5.5	t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisations of Cox and GBT model agreement in CKB individuals.	72
6.1	Inclusion criteria for analyses presented in Chapter 6.	80
6.2	Timing of baseline questionnaire, resurvey 1, and resurvey 2 for included CKB individuals.	86
6.3	Temporal trends in predicted 3-year stroke risk for a random sample of (a) 50 men and (b) 50 women from CKB, stratified by incident stroke outcomes.	88
6.4	Relationship between predicted 3-year stroke risk at second resurvey and per annum increase in predicted risk (fixed effect slope + random effect slope from linear mixed effects model) for (a) men and (b) women in CKB.	89
7.1	Overview of two modelling approaches for identifying individuals whose first stroke event is most likely to be haemorrhagic.	100
7.2	9-year probability of first incident stroke and first stroke death from first IS and first HS cases by age group with stratification by sex and type of area.	104
7.3	Cox risk estimates for 9-year risk of IS and HS in urban areas and rural areas.	107
7.4	Comparison of discrimination of individuals with first IS and first HS between Cox and GBT models.	107
7.5	Predicted number of events and deaths avoided for IS and HS per 10,000 patients treated by lowering LDL-C by 1 mmol l ⁻¹ with statins for 9-year follow-up period in Chinese adults with different levels of vascular risk.	110
7.6	Predicted number of events and deaths avoided for IS, HS, and MCEs per 10,000 patients treated by lowering LDL-C by 1 mmol l ⁻¹ with statins for 9-year follow-up period in Chinese adults with different levels of vascular risk.	112
8.1	t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisations of primary IS cases in CKB.	119
8.2	Flow chart of investigations and adjudicated subtypes of IS cases in CKB.	121
8.3	Proportions of IS subtypes by area in China.	122
8.4	Estimated cumulative all-cause mortality and recurrent stroke rates after first IS of different subtypes.	124
8.5	Performance of multiclass LR model on evident and probable cases of each IS subtype in test set.	126

8.6	Performance of multiclass LR model on possible cases of each IS subtype in test set.	126
8.7	All-cause mortality and recurrent stroke rates of different IS subtypes for (i) cases with evident and probable determined aetiology and (ii) ML-classified strokes with undetermined aetiology due to incomplete investigation.	128
B.1	Test of PH assumption in CKB men.	148
B.2	Test of PH assumption in CKB women.	149
B.3	Multivariable hazard ratios (HRs) and 95% CI for total stroke using Framingham (“2017 FSRP” models) versus CKB (“+Area Stratification” models, modified to include ever-regular smoking and age started smoking) coefficients in men and women.	149
B.4	9 year incidence for (A) IS and (B) HS by CKB area (adjusted for competing risk of death).	150
C.1	Grid search for MLP architecture selection.	158
C.2	Model loss and accuracy against epochs for best-performing and worst-performing MLP model architectures.	159
C.3	AUROC for prediction of 9 year risk of stroke.	162
C.4	Hosmer-Lemeshow/Nam-D’Agostino test statistics for prediction of 9 year risk of stroke.	162
C.5	Calibration plots by age group for the Cox and ML models in men and women.	163
D.1	Distribution of age at baseline, resurvey 1, and resurvey 2 in included CKB individuals.	168
D.2	Distribution of SBP at baseline, resurvey 1, and resurvey 2 in included CKB individuals.	169
D.3	Distribution of self-reported smoking status and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals.	169
D.4	Distribution of history of CHD and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals.	170
D.5	Distribution of diabetes status and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals.	171
D.6	Distribution of self-reported use of antihypertensive medication and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals.	172
D.7	Calibration plots for “Cox: Single measurement at most recent visit” models in men and women.	172

D.8	Calibration plots for “GBT: Single measurement at most recent visit” models in men and women.	173
D.9	Calibration plots for “GBT: Sequential measurements at three visits” models in men and women.	173
D.10	Calibration plots for “GBT: Longitudinal summary of sequential measurements at three visits” models in men and women.	173
D.11	Calibration plots for “Longitudinal summary of stroke risk estimates at three visits + Single measurement at most recent visit” models in men and women.	174
E.1	9-year incidence and incident mortality from first ischaemic stroke (IS) and first haemorrhagic stroke (HS) cases by age group with stratification by vascular risk.	176
E.2	Discrimination and calibration performance of Cox models for predicting nine-year absolute risks of ischaemic stroke (IS).	177
E.3	Discrimination and calibration performance of Cox models for predicting nine-year absolute risks of haemorrhagic stroke (HS).	177
E.4	9-year probability of incident major coronary event (MCE) and MCE death by age group with stratification by sex, type of area, and vascular risk.	179
F.1	Performance of multiclass LR model on evident and probable cases of each IS subtype in test set (excluding silent cerebral infarcts). . .	186
F.2	Performance of multiclass LR model on possible cases of each IS subtype in test set (excluding silent cerebral infarcts).	186
F.3	Estimated cumulative all-cause mortality and recurrent stroke rates after first IS of (i) different subtypes (excluding silent cerebral infarcts) and (ii) silent cerebral infarcts.	187
F.4	All-cause mortality and recurrent stroke rates of different IS subtypes for (i) cases with evident and probable determined aetiology and (ii) ML-classified strokes with undetermined aetiology due to incomplete investigation (excluding silent cerebral infarcts).	187

List of Tables

2.1	ICD-10 codes for cerebrovascular diseases.	9
2.2	Existing stroke risk prediction models.	15
2.3	Distribution of risk factors used in existing stroke risk prediction models.	17
4.1	Distribution of established risk factors for total stroke and stroke pathological types in men and women in CKB.	45
4.2	Comparison of performance of different models for prediction of total stroke and stroke pathological types in men and women in CKB. . .	47
5.1	Discrimination and calibration performance of Cox and ML models for prediction of 9-year risk of stroke.	68
5.2	Top 10 most important GBT features for risk prediction of stroke in men and women from 0-3 years, 3-6 years, and 6-9 years.	71
5.3	Top 10 most important decision tree features for selecting between Cox and GBT for ensemble risk screening approach.	73
5.4	Summary metrics from screening of high-risk individuals in test set using (i) a Cox-only approach, (ii) a GBT-only approach, and (iii) an ensemble approach.	74
5.5	Agreement among Cox-only, GBT-only, and ensemble approaches for screening of high-risk individuals in the test set.	74
6.1	Distribution of major risk factors for stroke in individuals at baseline (2004-2008), 1st resurvey (2008), and 2nd resurvey (2013-2014) in CKB.	87
6.2	Comparison of discrimination and calibration performance for prediction of 3-year risk of stroke using Cox and GBT modelling approaches. . .	90
6.3	Δ AUC, NRI and IDI of Cox and GBT modelling approaches for predicting 3-year risk of stroke.	91
6.4	Top 10 most important features for GBT models in men.	92
6.5	Top 10 most important features for GBT models in women.	93
7.1	Baseline characteristics of included participants from CKB, stratified by pathological type of first stroke event.	103

7.2	Top 10 important features by effect size in Cox models for IS in men and women.	106
7.3	Top 10 important features by effect size in Cox models for HS in men and women.	106
7.4	Top 20 important features used by GBT models for discriminating between a first IS and first HS event.	108
7.5	Numbers and case fatality of reported first events of IS, HS, and MCE.	111
8.1	Key variables collected from medical records for classification of IS subtypes using CCS.	118
8.2	Baseline characteristics by IS subtypes.	123
8.3	Breakdown of second strokes by pathological types following a first stroke of different IS subtypes.	125
8.4	Top risk factors for classifying IS subtypes.	127
A.1	CKB risk factor indicators.	142
A.2	CKB stroke adjudication variables.	146
B.1	Area-specific discrimination performance for total stroke and stroke pathological types in men and women using “+Area Stratification” models with FSRP inputs.	150
B.2	Area-specific discrimination performance for total stroke and stroke pathological types in men and women using “+Additional Risk Factors” models.	150
B.3	Discrimination and calibration performance for total stroke and stroke pathological types in men and women (age 55-84) using 2017 Framingham Stroke Risk Profile.	151
B.4	Discrimination and calibration performance for total stroke and stroke pathological types in men and women (age 55-84) using 1991 Framingham Stroke Risk Equation (Wolf).	152
B.5	Discrimination and calibration performance for total stroke and stroke pathological types in men and women (age 55-84) using 1994 Framingham Stroke Risk Equation (D’Agostino).	153
B.6	Comparison of performance of different models for prediction of total stroke and stroke pathological types in men and women (with modified ordering of incremental changes compared to Table 4.2).	154
C.1	Optimal parameter values and validation AUROCs for various MLP regularisation techniques.	160
C.2	Discrimination and calibration performance of Cox and ML models for risk prediction of stroke from 0-3 years.	163

C.3	Discrimination and calibration performance of Cox and ML models for risk prediction of stroke from 3-6 years.	164
C.4	Discrimination and calibration performance of Cox and ML models for risk prediction of stroke from 6-9 years.	164
C.5	Confusion matrices for Cox-only, GBT-only, and ensemble approaches for screening of high-risk individuals in the test set.	165
E.1	Baseline characteristics of included participants from the China Kadoorie Biobank with an incident major coronary event (MCE) prior to any stroke event.	178
F.1	Proportions of IS subtypes in the training set and test set.	183
F.2	Breakdown of primary IS cases with separation of silent and non-silent strokes.	184
F.3	Top risk factors for classifying IS subtypes (excluding silent cerebral infarcts).	185

List of Abbreviations

AF		Atrial fibrillation
AHA		American Heart Association
ARIC		Atherosclerosis Risk in Communities Study
ASCVD		Atherosclerotic cardiovascular disease
ASE		Anderson Stroke Equation
AUC/AUROC		Area under the receiver operating characteristic curve
BMI		Body mass index
BP		Blood pressure
CART		Classification and Regression Trees
CCS		Causative Classification System for Ischemic Stroke
CE		Cardioaortic embolism
CHD		Coronary heart disease
CHS		Cardiovascular Health Study
CI		Confidence interval
CKB		China Kadoorie Biobank
CT		Computerised tomography
CVD		Cardiovascular disease
DBP		Disatolic blood pressure
DNN		Deep neural network
DALY		Disability-adjusted life year
EC		Extracranial
ECG		Electrocardiogram
Echo		Echocardiography
EHR		Electronic health record
ESRS		Essen Stroke Risk Score

FSRS		Framingham Stroke Risk Score
FSRP		2017 Framingham Stroke Risk Profile
GBT		Gradient boosted tree
HDL		High-density lipoprotein
HIC		High-income country
HR		Hazard ratio
HS		Haemorrhagic stroke
IC		Intracranial
ICD-10		International Classification of Diseases 10 th Revision
ICD-11		International Classification of Diseases 11 th Revision
ICH		Intracerebral haemorrhage
IDI		Integrated discrimination improvement
IHD		Ischaemic heart disease
IS		Ischaemic stroke
KSRP		Korea Stroke Risk Profile
LAA		Large artery atherosclerosis
LACI		Lacunar stroke
LASSO		Least absolute shrinkage and selection operator
LDL		Low-density lipoprotein
LME		Linear mixed effects model
LMIC		Low- and middle- income country
LR		Logistic regression
MCE		Major coronary event
MI		Myocardial infarction
ML		Machine learning
MLP		Multilayer perceptron
MRI		Magnetic resonance imaging
NHANES III		Third National Health and Nutrition Examination Survey
NHIRD		Taiwanese National Health Insurance Research Database
NPV		Negative predictive value
NRI		Net reclassification index

OR		Odds ratio
PCE		Pooled Cohort Equations
PH		Proportional hazards
PPV		Positive predictive value
QALY		Quality-adjusted life year
RSF		Random survival forest
SAH		Subarachnoid haemorrhage
SAO		Small artery occlusion
SBP		Systolic blood pressure
SD		Standard deviation
SVM		Support vector machine
TIA		Transient ischaemic attack
TOAST		Trial of Org 10172 in Acute Stroke Treatment
t-SNE		t-distributed stochastic neighbor embedding
UKPDS		United Kingdom Prospective Diabetes Study

Prevention is better than cure.

— Desiderius Erasmus

1

Introduction

1.1 Introduction to Stroke in China

Cardiovascular disease (CVD), including ischaemic heart disease (IHD) and stroke, has nearly doubled in global prevalence from 1990-2019 and is the leading cause of annual mortality, accounting for 18.6 million deaths in 2019 (32.8% of all deaths) [1]. Stroke, in particular, was the cause of 6.6 million deaths worldwide and 5.7% of global disability-adjusted life years (DALYs) from all causes [2]. While global stroke incidence has decreased in most regions (except East Asia and Sub-Saharan Africa) since 1990, the absolute number of people affected by stroke has increased for both men and women of all ages, highlighting the need for improvements in stroke risk prediction and prevention [3]. By preemptively modifying levels of established risk factors, it is estimated that over half of all strokes may be prevented [4].

In order to address the enormous global burden of stroke, there is still a need to understand the epidemiology and treatment of stroke in different settings of the world. Despite the fact that 75% of stroke deaths and 81% of stroke-related DALYs occur in low- and middle-income countries (LMICs), most current guidelines for stroke management are based on evidence derived from high-income countries (HICs) [5]. Consequently, there is substantial uncertainty regarding best practices in LMICs, where stroke is marked by increasing incidence, earlier age of onset, and

higher proportion of haemorrhagic stroke (HS) types compared to HICs [6]. Better models for risk prediction and improved understanding of the roles of major risk factors in these regions are essential to reducing the burden of stroke worldwide.

China is a region of primary interest because it bears the largest stroke burden in the world. Stroke is the leading cause of death in China, which accounts for nearly one-third of stroke deaths worldwide [2]. Moreover, stroke incidence has increased in China over the past 30 years, especially in rural areas [7]. As in other LMICs, the proportion of HS in China is significantly higher than in HICs, highlighting the need to develop new risk prediction models for stroke tailored to a contemporary Chinese population.

In addition to having large geographic variation, stroke is also a heterogeneous disease having various pathologies and aetiologies with different risk factors and outcomes. Most risk prediction models for stroke [8–10] evaluate risk of ischaemic stroke (IS) only, or aggregate IS and HS events into a single “total stroke” endpoint. Limited work has been done on separately evaluating risk for each stroke pathological type [11, 12], but such models may have significant implications for optimising primary prevention strategies and increasing the effective use of statin therapy in China.

Moreover, even among IS, strokes can be further divided into heterogeneous aetiological subtypes including cardioaortic embolism (CE), large artery atherosclerosis (LAA), and small artery occlusion (SAO). Existing algorithms for classifying IS subtypes such as the Trial of Org 10172 in Acute Stroke Treatment (TOAST) [13] or Causative Classification System for Ischemic Stroke (CCS) [14] often label substantial proportions of IS events as having undetermined aetiology due to lack of evidence from brain imaging, vascular tests, and cardiac tests. Consequently, there is considerable potential for using risk models to provide probabilistic classification of IS subtypes in the absence of such evidence. Accurate subtyping of additional IS events can help guide treatment decisions (e.g., the use of antihypertensive medication) for affected individuals as well as increase the statistical power of genetic studies that will broaden our scientific understanding of stroke aetiology.

In summary, the primary prevention of stroke is a high-priority clinical challenge at the forefront of global health. Building the tools and insight to address this clinical challenge entails the application of engineering and data science principles at multiple stages including (i) the development and validation of appropriate risk prediction models in LMICs such as China, (ii) optimising the use of preventative measures informed by such models, and (iii) exploring the use of risk factor data to yield probabilistic classifications of IS subtypes that would otherwise be left unclassified by existing classification algorithms.

1.2 Thesis Objectives and Overview

The aims of this thesis are to apply a combination of traditional statistical methodologies and modern machine learning (ML) approaches to Chinese healthcare data to (i) develop risk prediction algorithms that elucidate the joint effects of multiple factors on total stroke and different pathological stroke types, (ii) optimise the use of these algorithms to minimise stroke events and fatalities in China, and (iii) better understand the heterogeneity of IS subtypes to provide probabilistic classifications in real-world settings with incomplete evaluation data. The contributions of this thesis are presented with the following structure:

- Chapter 2 provides a review of stroke pathology, current guidelines on primary prevention of stroke, and existing stroke risk prediction models.
- Chapter 3 describes the China Kadoorie Biobank (CKB) – a prospective cohort study of 0.5 million contemporary Chinese adults used as the primary data source for this thesis.
- Chapter 4 analyses the performance of the 2017 Framingham Stroke Risk Profile (FSRP) in CKB. We then compare the FSRP to novel CKB risk scores for total stroke, IS, and HS developed using a traditional Cox regression approach.

- Chapter 5 describes the development and evaluation of various ML approaches for risk prediction of total stroke as well as an ensemble risk screening approach for identifying individuals at high-risk of stroke.
- Chapter 6 compares the utility of single versus sequential risk factor measurements for risk prediction of stroke using a combination of linear mixed effects models and gradient boosted trees to model temporal trends in stroke risk.
- Chapter 7 describes the development and evaluation of prediction models that assess differential risk of a first-ever stroke event being an IS or HS event. We then estimate the stroke events and fatalities that could be avoided by optimising the use of statin therapy as informed by these models. We further consider the effects of statins on major coronary events (MCEs) for informing preventative medication strategies in China.
- Chapter 8 investigates the heterogeneity of risk factors and outcomes for IS subtypes. We also explore the use of a multiclass, regularised logistic regression (LR) approach for providing probabilistic classification of IS events with undetermined aetiology according to CCS.
- Chapter 9 provides a summary of the thesis and discusses directions for future work.

2

Background

2.1 Pathology of Stroke

2.1.1 Stroke Definitions

Since 1970, stroke has been defined by the World Health Organization as “rapidly developed clinical signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death, with no apparent cause other than of vascular origin [15].” Sudden, focal neurological deficits of vascular origin lasting less than 24 hours are generally differentiated from stroke and classified as transient ischaemic attacks (TIAs).

Although the World Health Organization’s definition of stroke is still widely used, it is not universally accepted due to its heavy emphasis on clinical symptoms [16]. For example, in 2013, the American Heart Association (AHA) revised their endorsed definition of stroke to include silent infarction and silent haemorrhage unaccompanied by classically defined stroke syndromes [17]. This change moved towards a structural, rather than solely clinical, diagnosis of stroke, and may have significant impact on the recording of strokes in different parts of the world. As of today, no consensus has been reached and reporting remains unstandardised.

2.1.2 Stroke Pathological Types

Regardless of definition, stroke can be broadly classified into two main categories or pathological types: ischaemic and haemorrhagic stroke (Figure 2.1). Ischaemic strokes (IS) occur when a blood vessel supplying blood to the brain is obstructed, resulting in cerebral infarction. Meanwhile, haemorrhagic strokes (HS) are caused by leakage or rupture of such a blood vessel [18]. Globally, IS has higher prevalence than HS, accounting for 76% of stroke prevalence in 2019 [2]. However, HS is much more fatal than IS, accounting for half of the 6.6 million global stroke deaths during the same period. Moreover, the proportion of HS in low- and middle-income countries (LMICs) such as China, is significantly greater than in high-income countries (HICs), especially in rural areas [7]. Although the causes of these differences have not been fully explained, it is hypothesised that differences in quality of stroke primary prevention, level of stroke awareness, and management of hypertension may all be contributing factors.

In addition to having distinct outcomes, IS and HS are also known to have differing risk factors and should be managed differently. For example, while increased adiposity and LDL-C levels are risk factors for IS, they can have a protective effect for HS [11, 19, 20]. Meanwhile elevated blood pressure has been found to be associated with a higher risk of HS than IS [21, 22]. Moreover, while anti-platelet drugs and cholesterol-lowering medication such as statins are currently recommended for IS patients, they are not recommended for individuals with HS [23]. Further study of IS and HS as separate disease endpoints is crucial to advancing our understanding of these heterogeneous stroke pathologies in different populations around the world.

2.1.3 Subtypes of Ischaemic Stroke

Ischaemic stroke (IS) can be further subtyped into distinct aetiological categories. The three main causative subtypes of IS are cardioaortic embolism (CE), large artery atherosclerosis (LAA), and small artery occlusion (SAO) (Figure 2.1). CE is caused by arterial occlusions due to an embolus arising in the heart, most commonly resulting from non-valvular atrial fibrillation. LAA is caused by atherosclerosis

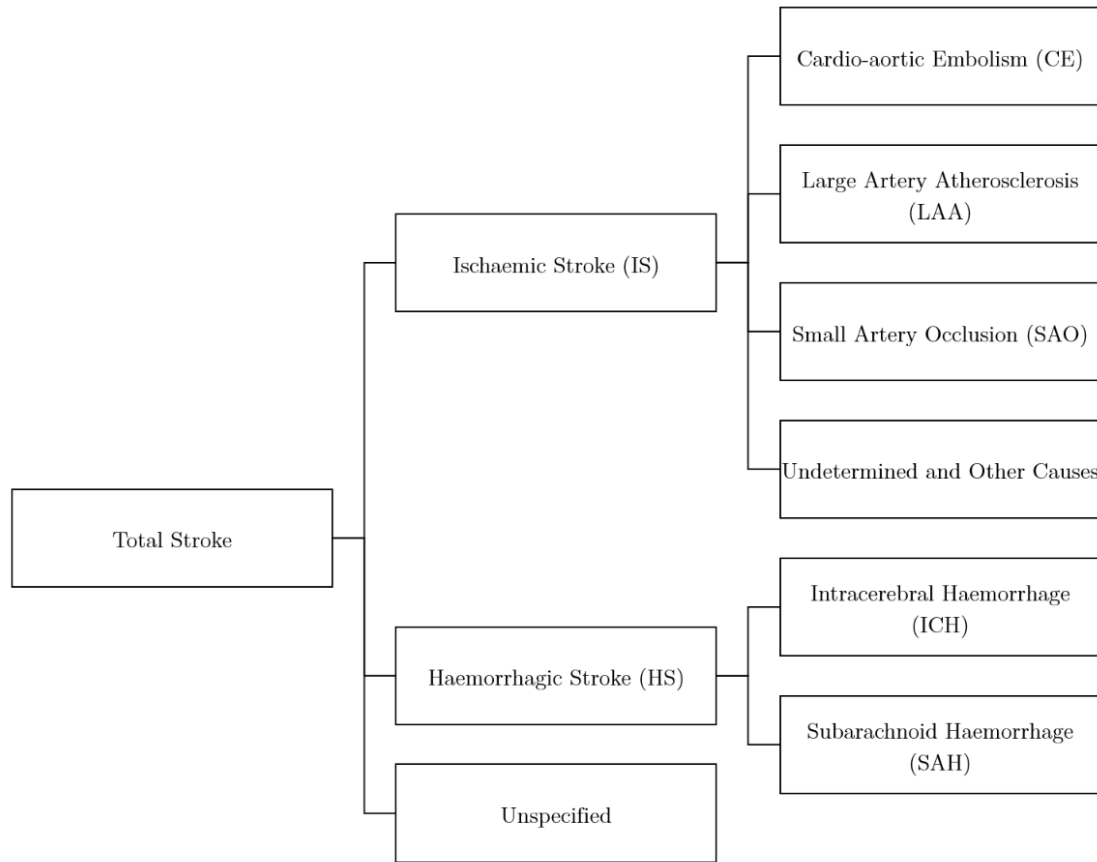


Figure 2.1: Pathological types and subtypes of stroke.

of the major vessels supplying the brain (e.g., extracranial and large intracranial arteries), which may cause ischaemia through embolism or haemodynamic causes. SAO, sometimes referred to as lacunar stroke, is caused by small sub-cortical infarcts (<15-20mm in diameter) resulting from occlusion of a single penetrating artery. When symptomatic, SAOs are typically accompanied by a classical “lacunar” syndrome such as pure motor hemiparesis, pure sensory stroke, sensorimotor stroke, dysarthria-clumsy hand syndrome, and ataxic hemiparesis [13, 14, 18].

The most widely used current clinical adjudication algorithm for IS subtyping is the Trial of Org 10172 in Acute Stroke Treatment (TOAST) [13], favoured for its ease of use and good inter-observer agreement [18]. In addition to the three main pathogenic mechanisms listed above, TOAST includes categories for (i) other rare determined causes (e.g., nonatherosclerotic vasculopathies, hypercoagulable states, or hematologic disorders) and (ii) undetermined causes. IS strokes des-

ignated by TOAST as having an “undetermined” cause comprise strokes that are “undetermined” because of incomplete investigation and strokes that are fully investigated, but cryptogenic.

Another commonly used adjudication algorithm is the Causative Classification System for Ischemic Stroke (CCS) [14], which is a computerised version of a modified TOAST algorithm called SSS-TOAST [24]. Compared to TOAST, CCS requires additional steps for implementation but provides higher inter-examiner reliability. CCS also includes a further refinement of each IS subtype to account for advances in diagnostic tools, and provides evident, probable, and possible classifications for each IS subtype based on the level of supporting clinical, diagnostic, and epidemiological evidence.

Finally, in cases where a singular aetiological subtype classification is not required, a phenotypic classification called ASCO/ASCOD can be used, in which stroke subtypes are assigned into non-mutually exclusive categories without weighting toward the most likely cause in the presence of competing aetiologies [25, 26].

Regardless of the classification scheme used, it is well known that IS subtypes differ in their risk factors and outcomes, and should inform stroke management accordingly. For example, SAOs are often “silent” strokes, producing no clinically recognised acute stroke symptoms, and have lower mortality rates than LAA. Meanwhile, individuals with CE are at the highest risk for recurrent strokes and 1-month mortality [13, 18]. Similar to the distribution of IS and HS, evidence from a systematic review suggests that the distribution of IS subtypes also differs between Chinese and white populations [27], further motivating the study of these heterogeneous subtypes in China.

2.1.4 Subtypes of Haemorrhagic Stroke

Haemorrhagic strokes (HS) can be subtyped into either intracerebral haemorrhage (ICH) or subarachnoid haemorrhage (SAH), depending on the location of the bleed (Figure 2.1). ICH is characterised by bleeding from an arterial source directly into the brain parenchyma, predominantly in the deep portions of the cerebral hemispheres

[18]. On the other hand, SAH is characterised by bleeding into the space between the arachnoid mater and pia mater of the meninges, located just outside of the brain.

Since HS subtypes are classified based on location of the bleed, subtyping can largely be performed using brain imaging alone. This is in contrast to IS subtyping, which is based on aetiology and involves a combination of evidence from clinical features, imaging, and medical tests.

Globally, there were 3.4 million incident cases of ICH in 2019 (28% of total stroke incidence), while SAH accounted for 1.2 million incident cases (10% of total stroke incidence). However, ICH had much higher mortality, accounting for 44% of global stroke deaths, while SAH accounted for only 6% of stroke mortality [2]. In China, ICH comprises 27%-51% of all strokes (compared to 6%-20% in predominantly white European populations) [27], and it is ICH, not SAH, that drives the difference in proportion of HS between these populations.

2.1.5 ICD-10 Codes for Stroke Pathologies

The International Classification of Diseases 10th Revision (ICD-10) is an “international standard diagnostic classification for all general epidemiological and many health management purposes,” and allows for systematic recording, comparison, and interpretation of data from various sources [28]. A range of heterogeneous cerebrovascular diseases are covered in ICD-10 by the codes I60 - I69 (Table 2.1).

Table 2.1: ICD-10 codes for cerebrovascular diseases.

I60	Subarachnoid haemorrhage (SAH)
I61	Intracerebral haemorrhage (ICH)
I62	Other nontraumatic intracranial haemorrhage
I63	Cerebral infarction
I64	Stroke, not specified as haemorrhage or infarction
I65	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
I66	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
I67	Other cerebrovascular diseases
I68	Cerebrovascular disorders in diseases classified elsewhere
I69	Sequelae of cerebrovascular disease

Mapping the terminology used throughout this thesis to standard ICD-10 codes, we use the term “ischaemic stroke (IS)” to refer to any event coded as I63. The term “haemorrhagic stroke (HS)” is used to refer to any event coded as either I60 (SAH) or I61 (ICH). Lastly, the term “unspecified” stroke refers to any event coded as I64. For IS subtypes, we do not refer to the diagnostic criteria of ICD-10, but rather utilise the adjudication criteria specified by the Causative Classification System for Ischemic Stroke (CCS) [14], except where otherwise noted.

2.2 Primary Prevention of Stroke

Stroke may be a leading cause of death and disability, but it is also a preventable disease. Through a focus on primary prevention (i.e., prevention of stroke before a first event ever occurs), it is estimated that 5-year global stroke incidence can be reduced by 50% [4]. Primary prevention is especially crucial because over three-quarters of strokes are first events [29]. Effective primary prevention of stroke relies on successful modification of established stroke risk factors using a combination of population-based strategies and risk-based approaches that target individuals with the highest risk of stroke. Such strategies are particularly important in LMICs such as China, which now bear the brunt of the global stroke burden, but have lagging health care services and limited resources for stroke prevention [5]. In this section, we review the available evidence on stroke risk factors, proposed strategies for primary prevention, and recommendations provided by current clinical guidelines.

2.2.1 Risk Factors

Current guidelines on primary prevention of stroke from the American Heart Association (AHA) [30] and the Chinese Stroke Association [31] are in alignment and classify risk factors into three categories: (i) “nonmodifiable risk factors,” (ii) “well-documented and modifiable risk factors,” and “less well-documented or potentially modifiable risk factors.”

According to these guidelines, non-modifiable risk factors include increased age, low birth weight, race/ethnicity, and genetic factors. While identification of

these non-modifiable factors can be useful for reliable risk assessment, there is limited evidence for the ability to modify stroke risk attributable to these factors, and genetic screening of the general population is currently not recommended for prevention of a first stroke.

Well-documented modifiable risk factors are those with strong “clear, supportive epidemiological evidence and evidence of risk reduction when modified in the context of randomized clinical trials [30].” These include physical inactivity, dyslipidaemia, diet and nutrition, hypertension, obesity and body fat distribution, diabetes mellitus (particularly type 2 diabetes), cigarette smoking, atrial fibrillation (AF), other cardiac conditions (e.g., myocardial infarction [MI], cardiomyopathy, valvular heart disease, etc.), asymptomatic carotid stenosis, and sickle cell disease. With respect to dyslipidaemia, low levels of total cholesterol (and especially LDL cholesterol) have been associated with increased risk of ICH, while high levels have been associated with increased risk of IS. There is also evidence for an inverse relation of HDL cholesterol and risk of stroke. With respect to diet and nutrition, high levels of salt intake, low levels of potassium intake, excessive weight, high alcohol consumption, and suboptimal dietary patterns all increase risk of stroke, while a diet rich in fruits and vegetables (e.g, the Mediterranean diet) has been associated with decreased risk. Less well-documented modifiable risk factors of stroke include migraine, metabolic syndrome, light to moderate alcohol consumption, drug abuse, sleep-disordered breathing, hyperhomocysteinaemia, elevated Lp(a), hypercoagulability, inflammation, and infection.

Importantly, INTERSTROKE, a case-control study including 22 countries found that five potentially modifiable risk factors are associated with 80% of the risk of stroke: hypertension, current smoking, abdominal obesity, diet, and physical activity [32]. Meanwhile, an extra 10% of stroke risk can be attributed to the additional consideration of diabetes, alcohol intake, psychosocial factors (stress and depression), cardiac causes (AF, previous MI, rheumatic valvular disease, and prosthetic heart valve), and concentrations of apolipoproteins. This suggests that

there is a opportunity to significantly reduce strokes by proactively modifying a handful of key risk factors.

The major risk factors are also mostly similar in HICs and LMICs, apart from some differences in effect size [32]. LMICs do have additional risk factors including infectious diseases such as tuberculosis, syphilis, HIV, malaria, etc., particularly in Africa where these conditions are more prevalent [5]. There are also differences in the distributions of some risk factors. For example, China accounts for almost one-third of the world’s smokers, with 52.1% of men being current smokers, but only 2.7% of women in 2015 [33]. However, as a whole, primary prevention targets across international settings should be largely similar, as supported by the alignment of recent Chinese and American stroke prevention guidelines [30, 31].

2.2.2 Population-Based Strategies

Population-based approaches for primary prevention of stroke target the entire population to modify risk factors and reduce stroke risk. These include mass health promotion strategies that utilise media, legislative measures, and health education to decrease smoking, increase exercise, and improve diets. Some proposals have also advocated the use of mobile technologies to reach large portions of the population [5, 34].

The AHA’s guidelines for primary prevention of stroke [30] propose that emergency departments hold much potential to provide health promotion and disease prevention services such as the screening and management of hypertension, diabetes, smoking behaviour, AF, and alcohol consumption. Moreover, the guidelines emphasise the importance of improving preventative vascular services from primary care physicians and specialists through modifying the organisation or context in which care is delivered and implementing stroke education programs. In LMICS, further emphasis has been placed on the role of community health workers in increasing recruitment and improving adherence to lifestyle changes and preventative medications [35].

In China, public education and primary prevention initiatives for stroke have been implemented [36–38], and from 2002-2012 awareness rate, treatment rate, and control rate of hypertension and diabetes have all increased substantially along with a decreased use of tobacco [39]. In 2012, the Chinese Center for Disease Control and Prevention drafted a China National Plan for Non-Communicable Diseases Prevention and Treatment to further increase awareness, improve control of hypertension and diabetes, lower smoking and salt intake, and enhance disease surveillance [40]. Since then, the Chinese government has also sponsored several regional and nationwide studies to identify gaps between clinical practice and recommended guidelines, and to design and evaluate intervention tools for improving quality of stroke care and patient outcomes [41–43]. Despite these advances, there remains significant room for improvement of stroke prevention in China and other LMICs, and continual progress is needed.

2.2.3 Risk-Based Strategies

Risk-based strategies for primary prevention of stroke seek to identify individuals with the highest risk of developing stroke and targeting preventive measures, including lifestyle changes and pharmacological interventions, among such individuals. Lifestyle changes should be considered as the first means for stroke prevention. Among the most important changes are improved diet, increased physical activity, weight loss, smoking cessation, and management of psychosocial stress [5]. In combination with lifestyle changes, pharmacological interventions may also be considered. For example, guidelines recommend the use of statins for individuals with high cholesterol, antihypertensive drugs for blood pressure management, and antithrombotic treatment for patients with AF and other cardiac conditions.

One proposed approach to pharmacological intervention includes the use of a polypill, or a single tablet containing a combination of lipid-lowering drugs, blood pressure-lowering drugs, and sometimes aspirin [4, 44]. This may be particularly attractive for LMICs with a relatively young population at high-risk, low awareness of risk factors, and low compliance in primary and secondary prevention treatment.

With respect to implementation, it has been suggested that such a polypill could be utilised in a risk-based mass-strategy approach. For example, the polypill can be recommended for wide portions of the population with the highest absolute risk of CVD and stroke, or in individuals with the highest risk relative to others of the same age and sex [4, 34]. It is believed that the polypill might increase compliance, especially if implemented in a mass-strategy approach, but further evidence is required from randomised clinical trials.

In China, multicentre trials combining both population-based strategies and risk-based strategies have been implemented with success, suggesting that community-based intervention is feasible and effective for primary prevention of stroke [45].

2.3 Stroke Risk Models

In order to implement the risk-based strategies for stroke prevention described above, it is important to be able to accurately evaluate the risk of stroke at the individual level. Risk prediction tools, therefore, play an important role in reducing strokes. Moreover, because of the particular importance of primary prevention, risk prediction models are especially crucial for individuals who have never experienced a prior stroke or TIA event.

In practice, the AHA and American Stroke Association have emphasised the importance of including both stroke and coronary heart disease (CHD) events as outcomes in risk prediction tools for primary prevention [30]. Some of the most prominent CVD risk scores currently used around the world include 10-year atherosclerotic cardiovascular disease (ASCVD) risk as predicted by the 2013 Pooled Cohort Equations (PCE) [46], 10-year CVD risk as predicted by QRISK3 [47], and 10-year ASCVD risk as predicted by the China-PAR project [48]. Nevertheless, because of the distinct pathology of stroke, stroke-specific risk prediction models are still valuable, and current AHA guidelines suggest the use of stroke risk assessment tools to “identify people at elevated risk who might not be aware of their risk, to assess risk in the presence of >1 condition, to measure an individual’s risk that can be tracked and lowered by appropriate modifications, to estimate risk for

selecting treatments or stratification in clinical trials, and to guide appropriate use of further diagnostic testing [30].”

This thesis focuses specifically on stroke-specific risk evaluation, and in this section, we provide a review of existing models developed for stroke risk prediction, including those most commonly used in clinical practice.

Table 2.2: Existing stroke risk prediction models.

Model	Year	Training Population	Sample Size	Output
Framingham Stroke Risk Models				
Anderson Stroke Equation (ASE) [49]	1991	Framingham, MA, USA; 30-74 yo; free of CVD (1968-1975)	5,573	4-12 yr risk of first total stroke
Framingham Stroke Risk Score (FSRS) [50]	1991	Framingham, MA, USA; 55-84 yo; free of stroke (1960s-80s)	5,734	10 yr risk of first total stroke
D’Agostino FSRS [51]	1994	Framingham, MA, USA; 55-84 yo; free of stroke (1960s-80s)	5,734	10 yr risk of first total stroke
Revised Framingham Stroke Risk Profile (FSRP) [8]	2017	Framingham, MA, USA; 55-84 yo; free of stroke (1990s-2010)	5,072	5 yr & 10 yr risk of first total stroke & first IS
Additional Stroke Risk Models				
EUROSTROKE Risk Score [52]	2002	Males from three separate cohort studies in Finland, the Netherlands, and the UK	698	Lifetime risk of future stroke (based on 7.3 yr average follow up time)
Atherosclerosis Risk in Communities (ARIC) [53]	2004	US adults from ARIC cohort; 45-64 yo; free of stroke (1987-2000)	~11,000	10 yr risk of first IS
Korea Stroke Risk Profile (KSRP) [54]	2008	Korean adults; 30-84 yo; free of cancer, liver disease, CVD, respiratory disease (1992-2005)	~612,000	10 yr risk of first total stroke
Taiwan Study [55]	2010	Ethnically Chinese; ≥ 35 yo; free of stroke (1990-2007)	3,513	10 yr risk of first total stroke
QStroke [9]	2013	451 general practices in England and Wales (QResearch database); 25-84 yo; free of stroke and TIA (1998-2012)	3,549,478	10 yr risk of first IS

Ferket <i>et al.</i> [11]	2014	African Americans & Caucasians from ARIC, Rotterdam, and CHS cohorts; 45+ yo; free of stroke, no anticoagulation medicine (1980s-2000s)	27,493	10 yr risk of first IS and ICH
NEW-STROKE [56]	2017	US NHANES III (1988-1990s)	NR	10 yr risk of first total stroke
China-PAR [10]	2019	Two Chinese cohorts; 35-74 yo; free of ASCVD (1990s-2000s)	21,320	10 yr & lifetime risk of first total stroke
Stroke Risk Models for Individuals with Specific Prior Diseases				
CHADS ₂ [57]	2001	Medicare beneficiaries across 7 US states; 65-95 yo; nonrheumatic AF	1,733	Stroke risk for individuals with AF (6-point scale)
United Kingdom Prospective Diabetes Study (UKPDS) [58]	2002	UKPDS hospital patients with newly diagnosed diabetes; 25-65 yo (1977-1991)	4,549	4-20 yr probability of stroke after diagnosis of diabetes
Essen Stroke Risk Score (ESRS) [59]	2005	Subset of international CAPRIE cohort with IS (1990s)	6,431	1 yr risk of recurrent stroke and combined CVD events (10-point scale)
ABCD ² [60]	2007	Combined cohort from Northern CA, USA and Oxfordshire, UK	1,916	Stroke risk for days following TIA (7-point scale)
CHA ₂ DS ₂ -VASc [61]	2009	Subset of Euro Heart Survey for AF cohort with AF (2000s)	1,084	Stroke risk for individuals with AF (9-point scale)
Machine Learning-Based Stroke Risk Models				
Khosla <i>et al.</i> [62]	2010	US-based CHS cohort; 65+ yo; free of stroke	4,988	5 yr risk of first total stroke
Letham <i>et al.</i> [63]	2015	MarketScan Medicaid Multi-State Database (MDCD); diagnosed with AF	12,586	1 yr risk of total stroke following AF diagnosis
Hung <i>et al.</i> [64, 65]	2017	Taiwanese NHIRD database; 0-99 yo; free of stroke (2000-2010s)	~756,000	3 yr & 5 yr risk of first IS
Li <i>et al.</i> [66]	2019	China national stroke screening data; 40+ yo; free of stroke (2017)	408,330	“High stroke risk” binary classification
Orfanoudaki <i>et al.</i> [67]	2020	Framingham, MA, USA; 40+ yo; free of stroke	~3,300	10 yr risk of first total stroke

NR: Not Reported.

Table 2.3: Distribution of risk factors used in existing stroke risk prediction models.

Risk Factor	Framingham Stroke Risk Models				Additional Stroke Risk Models								Models for Individuals with Specific Prior Diseases							
	ASE	FSRS	D'Agostino	FSRP	EUROSTROKE	ARIC	KSRP	Taiwan	QStroke	Ferket	NEW-STROKE	China-PAR	CHADS ²	UKPDS	ESRS	ABCD ²	CHA ₂ DS ₂ -VAS			
Age	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
Sex	X	X	X	X	*	X	X	X	X	X	X	X	X	X			X			
Blood Pressure	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
Antihypertensive therapy		X	X	X	X	X			X	X	X	X	X							
Diabetes	X	X	X	X	X	X	X	X	X	X	X	X	X	*	X	X	X			
Duration of diabetes												X		X						
Smoking	X	X	X	X	X	X	X		X	X	X			X						
Alcohol consumption						X	X													
Cardiovascular disease		X	X	X		X			X	X	X		X		X		X			
Atrial fibrillation		X	X	X				X	X		X		*	X			*			
Left ventricular hypertrophy	X	X	X			X					X									
Carotid artery wall thickness						X														
Peripheral arterial disease						X									X					
Previous TIA / ischaemic stroke					X								X		**	*	X			
Duration of TIA													X			X				
Speech disturbance																				
Unilateral weakness																X				
Physical exercise level							X		X		X					X				
Body mass index						X	X				X					X				
Waist circumference						X					X					X				
Waist:hip ratio										X										
Height											X									
Ethnicity						X			X	X	X									
Geographic area												X								
Urbanisation												X								
Deprivation score									X											
Rheumatoid arthritis									X											
Chronic renal disease									X											
Family medical history								X	X			X								
Hormone replacement therapy (postmenopausal females)											X									
Cholesterol levels	X					X	X	X	X	X	X	X					X			
Other biochemistry/lab risk factors						X		X		X										

*Excluded from risk factors because they are required qualifying events for use of the corresponding model.

**ESRS considers previous TIA or IS in addition to the required qualifying event for use of the model.

Note: ML-based risk models from Table 2.2 have been omitted here because their respective approaches often incorporate high numbers of clinically unestablished risk factors. Further discussion of these models, including the predictive features used, is provided in Section 2.3.2.

2.3.1 Traditional Approaches for Stroke Risk Prediction

Since the early 1990s, various risk prediction models have been developed to assess an individual’s absolute and relative risk of stroke based on clinically established risk factors from epidemiological studies. A representative sample of existing models, including those most commonly used in clinical practice, is shown in Table 2.2. The risk factors used in each model are provided in Table 2.3, with the exception of the ML-based approaches, which often include extensive numbers of inputs and are described in further detail in Section 2.3.2.

The Framingham Study, a multi-generation prospective cohort study including residents of Framingham, Massachusetts, USA, began in 1948 and has contributed extensively to our knowledge of the epidemiology of cardiovascular disease. Some of the earliest multivariable risk prediction models for stroke were developed using data from the Framingham cohort, and variants of such models are still widely considered to be the gold standard for assessing an individual’s probability of a first stroke event.

In 1990, Anderson *et al.* used data derived from the Framingham Study to develop a parametric, stroke-specific risk prediction model that imposed a Weibull distribution of survival times and output an individual’s probability of a first stroke for any time within 4 to 12 years of risk factor measurement [49]. Subsequently, the more widely recognised and commonly used Framingham Stroke Risk Score (FSRS) was developed based on a semi-parametric Cox proportional hazards approach, which has the benefit of making no assumptions of the underlying baseline hazard [50]. As with the Anderson Stroke Equation (ASE), the FSRS yielded an absolute measure of stroke risk, providing a probability of first-ever total stroke for any time after risk factor measurement. However, the FSRS is most commonly used for assessment of 10-year risk. Further variations of the FSRS have since been developed, including an update to adjust for the use of antihypertensive medication (“D’Agostino FSRS”) [51], and the 2017 Revised Framingham Stroke Risk Profile (FSRP) [8], which adjusted the FSRS for changing temporal trends in stroke and its related risk factors by using more recent cohort data collected between 1990 and 2010. These Framingham stroke risk models have been validated in both US and

European cohorts with robust discrimination [68–70], and hence, are recognised as the gold standard to predict stroke risk. However, a few population-based prospective studies have assessed the utility of these risk scores in Chinese adults, and although they have generally found modest discrimination of the Framingham models for total stroke, the models have been demonstrated to have poor prediction of absolute risk of stroke in these populations [10, 55, 71].

Additional risk models for stroke derived using traditional statistical approaches in other population-based studies include the 2002 EUROSTROKE Risk Score [52]; the 2004 Atherosclerosis Risk in Communities (ARIC) model [53]; the 2008 Korea Stroke Risk Profile (KSRP) [54]; a 2010 Taiwan study [55]; the 2013 QStroke model [9]; a 2014 study combining cohorts from the ARIC Study, Rotterdam Study, and Cardiovascular Health Study (CHS) [11]; the 2017 NEW-STROKE model [56]; and the 2019 China-PAR stroke models [10]. The EUROSTROKE Risk Score [52] was a logistic regression (LR) model based on three separate European cohorts from Finland, the Netherlands, and the United Kingdom and provided a lifetime risk score of having a future stroke event (based on data collected with an average follow-up time of 7.3 years). However, the model was limited to men only, and due to the use of LR, the risk score did not vary over the time domain as in the ASE or in the FSRs; only a single score was provided for each individual to characterise their lifetime risk of stroke. The ARIC model [53] used sex-specific Cox regression to derive 10-year risk scores for IS in a derivation cohort of ~11,000 individuals aged 45-64 from the USA recruited in the late 1980s. The KSRP [54] was a 10-year risk score for first total stroke developed using Cox models derived from a cohort of ~612,000 Koreans, aged 30-84 years who participated in at least one biennial medical examination offered by the National Health Insurance Corporation from 1992-1995 and were free of cancer, liver disease, CVD, and respiratory disease. The 2010 Taiwan study was conducted in a cohort of 3,513 ethnically Chinese adults (≥ 35 years old) who were stroke-free, yielding separate Cox models for predicting 10-year risk of total stroke using clinical versus biochemical features. QStroke [9] was another Cox model developed in a large cohort of over 3.5 million stroke-free

individuals living in England and Wales, and yielded 10-year risk estimates for first IS. Ferket *et al.* combined the ARIC, Rotterdam, and CHS cohorts to train Cox models on 27,493 individuals aged >45 years who had no prior stroke and did not use anticoagulation medication [11]. It is the only study to develop separate 10-year risk scores for IS and ICH, but it explicitly excluded any individuals who were not African American or Caucasian. The NEW-STROKE model [56], developed in 2017, was a proposed modification to the FSRs that added risk factors using data from the US Third National Health and Nutrition Examination Survey (NHANES III) cohort and was evaluated in the ARIC dataset. However, NEW-STROKE failed to report risk factor effect sizes as well as the sample size used from the US NHANES III cohort. Finally, the China-PAR stroke models [10] were derived from a pooled cohort of 21,320 ASCVD-free Chinese adults aged 35-74 years, who were recruited in the 1990s. The China-PAR stroke models included sex-specific Cox models for 10-year risk prediction of first total stroke and Fine-Gray competing risks models for predicting lifetime stroke risk.

A variety of additional stroke risk models have been designed for use in individuals with prior medical conditions. Some of the more frequently used of these models are included in Table 2.2 under “Stroke Risk Models for Individuals with Specific Prior Diseases.” For example, CHADS² [57], and its 2009 revision, CHA₂DS₂-VASc [61] are intended for risk stratification of individuals with AF. The United Kingdom Prospective Diabetes Study (UKPDS) risk model was designed for individuals with newly diagnosed diabetes [58]. The Essen Stroke Risk Score (ESRS) stratifies by one year risk of recurrent stroke and combined CVD events for individuals with prior stroke [59], and ABCD² stratifies by short-term stroke risk for individuals with a recent TIA [60]. While these models are helpful for clinical practice, they are not widely generalisable, leaving a need for a single cohesive model that can predict stroke risk for individuals with a variety of different comorbidities and medical histories.

Many of the traditional stroke risk models described in this section use very similar distributions of clinically established risk factors. Table 2.3 presents the inputs required by each model for stroke risk prediction/stratification and

demonstrates that virtually all models use age, sex, blood pressure, antihypertensive therapy, diabetes, smoking, and prior history of cardiovascular disease. Additional risk factors commonly included were ethnicity, body mass index (BMI), waist circumference, family medical history of CVD, and cholesterol levels. While the inclusion of additional risk factors can often improve the performance of risk prediction tools, these benefits must be weighed against practical considerations for cost-effective and widespread risk screening. For example, many people may not know the granular details of their medical condition (e.g., AF or left ventricular hypertrophy) or the levels of biochemical risk factors (e.g., cholesterol levels, fasting blood glucose, albumin, etc.), which are needed to implement more extensive models. This is especially true in LMICs where health education may lag and where diagnostic resources such as cardiac tests and blood tests are limited.

2.3.2 Machine Learning Approaches for Stroke Risk Prediction

With the increasing availability of high-performance computing and large health datasets, ML approaches have risen in popularity in recent years for a variety of healthcare applications, including the study of stroke. State-of-the-art ML techniques are able to learn from rich datasets and identify complex, non-linear interactions between risk factors, affording an opportunity to improve risk prediction algorithms for total stroke and stroke pathological types. Such improvements have already been observed for 10-yr risk prediction of cardiovascular disease (CVD), with ML approaches such as regularised logistic regression, random forest, gradient boosting, and neural networks demonstrating better risk discrimination, sensitivity, and specificity than the PCEs recommended in the American College of Cardiology guidelines [72].

Several ML models have also been developed for stroke-specific risk stratification and prediction. Khosla *et al.* demonstrated the use of an integrated ML approach including missing value imputation, automated feature selection, and learning algorithms (i.e., support vector machines [SVMs] and margin-based censored

regression) to improve the discrimination of 5-year stroke risk prediction over the traditional Cox modelling approach in the US-based CHS cohort [62]. Likewise, Hung *et al.* compared several ML approaches including a deep neural network (DNN), gradient boosted tree (GBT), LR, and SVM for 5-year stroke risk prediction using electronic medical claims from the Taiwanese National Health Insurance Research Database (NHIRD) and demonstrated that their DNN intelligent decision support system for IS risk assessment was competitive with risk models used in clinical practice for 3-year risk prediction [64, 65]. In another example, Letham *et al.*, used Bayesian Rule Lists, CART, C5.0, LR, SVM, random forest, and Bayesian CART to improve upon the discrimination yielded by the traditional CHADS₂ and CHA₂DS₂-VASc scores for 1-yr stroke risk prediction in individuals diagnosed with AF, using the US-based MarketScan Medicaid Multi-State Database (MDCD) [63]. Li *et al.* explored the use of LR, Naïve Bayes, Bayesian network, decision tree, DNN, random forest, bagging model, boosting model, and voting model algorithms to classify individuals as “high risk,” reporting good agreement with current risk classification methods used in China, even after artificially removing inputs that would cause existing screening approaches to fail [66]. Orfanoudaki *et al.* used longitudinal data from the Framingham Heart Study to increase their dataset size, and compared the FSRP with LR, CART, random forest, XGBoost, and the optimal classification trees algorithm [67]. They concluded that the improved performance of non-linear tree-based ML approaches over the FSRP suggested that there were important non-linear relationships between risk factors. However, it was unclear how much of this improvement was due to the inclusion of additional risk factors such as medication data, ECG results, and biochemical variables. While non-linear ML approaches did outperform regularised LR (a linear ML approach) trained with the same inputs, this improvement was much more modest compared to the reported improvements over the FSRP.

Compared to traditional stroke risk scores, which are often derived using Cox proportional hazards models, the ML approaches mentioned above have various differentiating characteristics. First, these ML models often employ a broader array

of input features such as extensive numbers of previous conditions and medications; various healthcare fees; and a range of metrics such as mini-mental score, predicted forced vital capacity, and ankle brachial index [62, 64]. In many cases, a large number of these inputs are not clinically established risk factors for stroke; nevertheless, they empirically contribute substantial value to predictive performance. This highlights the potential of ML approaches to identify previously unknown relationships in rich datasets to improve risk prediction and generate hypotheses that may motivate further clinical research. Moreover, compared to the traditional Cox model, in which non-linear interaction terms must be explicitly modeled, some ML methods such as tree-based algorithms, DNNs and non-linear SVMs are able to automatically identify non-linear relationships between risk factors, providing the potential for even better prediction of incident stroke [67]. Furthermore, while traditional stroke risk models make predictions solely based on a snapshot of an individual’s baseline characteristics at a particular time, some ML approaches also have the ability to handle longitudinal data to capture temporal risk effects. For example, Hung *et al.* engineered features from the longitudinal NHIRD data to capture mean values, standard deviations, and sums of risk factor values across several time periods of varying length [64]. Finally, some ML approaches, such as DNNs, have the ability to work with various data modalities (e.g., tabular data, free-text data, image data, etc.) as well as account for multiple competing outcomes [73, 74]. Thus ML models can potentially be developed and used for settings not currently addressed by traditional epidemiological methods.

As of today, ML models for stroke risk prediction have not yet been widely adopted in clinical practice and face several challenges. First, there is the issue of reproducibility. Validation studies of risk prediction tools in independent populations are a crucial step to becoming adopted into clinical guidelines. However, ML for health research performs notoriously poorly on several reproducibility metrics – even when using the same exact data [75]. Results are even less likely to be replicable across other populations and datasets, and in some cases, a number of the inputs required by a model may not be available at all in other settings. There

are also the issues of clinician understanding, explainability, and trust in ML risk prediction tools. Clinicians are much more likely to accept the decision of a software tool that they understand well enough to explain to a patient [76]. Consequently, complicated stroke risk models are unlikely to replace highly interpretable and established Cox models such as the FSRP in the clinical setting unless there is (i) strong evidence of improved performance and (ii) substantial efforts to explain the decisions of these models.

Despite these challenges, ML remains an important tool for increasing our knowledge of stroke epidemiology. Through increased sharing of anonymised health data and code for ML algorithms, application of rigorous statistical methods, and adherence to reporting guidelines such as the TRIPOD statement [77], issues of reproducibility in this field can be mitigated. Moreover, as electronic health records become more detailed and widespread, detailed individual-level data will be made increasingly available and enable the automatic updating and recalibration of complex ML models to local practices in various international settings. The development of better risk prediction models, the discovery of new risk factors, and the identification of complex risk factor interactions using ML can all serve to improve public health at a global scale.

2.4 Conclusion

As discussed in this chapter, stroke is a heterogeneous disease with multiple pathologies and appears to have different patterns in various regions of the world, including China. Through modification of key risk factors, primary prevention of stroke can contribute to reducing strokes by 50% globally, and an important step in this process is the development of risk prediction tools to identify individuals who have elevated risk. Traditional risk prediction models for stroke are typically Cox regression models with well established risk factor inputs, while more recent ML approaches using larger numbers of inputs have also been explored with promising results. However, existing risk prediction models have largely been derived in populations from HICs, with limited validation in LMICs such as China – especially

among more contemporary cohorts. Moreover, virtually all stroke risk models have focused on predicting total stroke or IS without separate prediction of HS, which is more fatal than IS and more common in developing countries. In the next several chapters, we will describe the work we have done to fill this knowledge gap by developing both traditional and ML-based approaches for evaluating risk of stroke and its various pathologies in a large contemporary population of 0.5 million Chinese adults.

3

The China Kadoorie Biobank

The data used throughout this thesis come from the China Kadoorie Biobank (CKB), a large and contemporary prospective cohort study of 0.5 million adults in China [78, 79]. In this chapter, we provide an overview of the CKB study design, discuss the validation and adjudication of stroke endpoints in CKB, and review the selected CKB data used in this thesis. We then provide key descriptive statistics of the CKB, including distributions of the most well-established risk factors for stroke as well as an analysis of missing values and censoring in the dataset.

3.1 Study Design

Designed to study the causes of common chronic diseases in China, the CKB included 512,891 adults aged 30-79 years. Data were collected using questionnaires, physical measurements, and blood samples collected for long-term storage. In order to account for regional variation, participants were recruited from 10 geographically defined areas (5 urban and 5 rural) in China based on local disease patterns, exposure to certain risk factors, population stability, quality of death and disease registries, local commitment, and capacity. The regions selected and numbers of participants recruited in each region are shown in Figure 3.1.

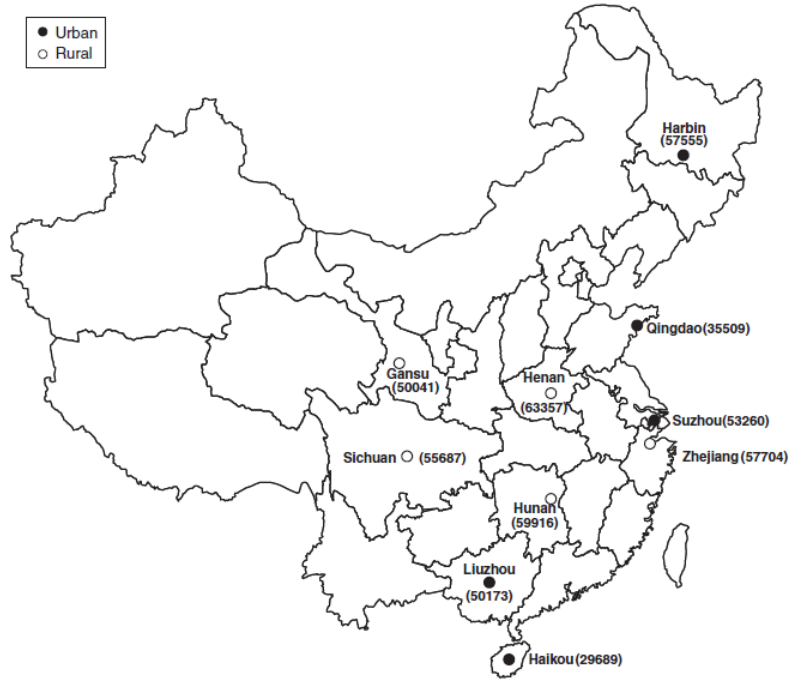


Figure 3.1: Locations of the 10 CKB survey sites and number of participants recruited. Open circles indicate rural areas and solid circles indicate urban areas. Number recruited at baseline in each area is shown in brackets. Source: Chen *et al.*, 2011 [79].

The CKB baseline questionnaire was administered between 2004-2008 and included questions about demographics, tea and alcohol consumption, smoking history, diet, pollution exposure, personal and family medical history, physical activity, reproductive history, sleeping habits, mood, and mental state. In addition, for each individual, a blood sample was collected and clinical examination included measuring standing height, sitting height, weight, hip circumference, waist circumference, bioimpedance, forced expiratory volume in 1 second, forced vital capacity, CO level, pulse rate, blood glucose, and blood pressure.

Following completion of the baseline survey, a first resurvey was completed from August to October 2008 involving 5% of randomly chosen survivors, with a second resurvey conducted from August 2013 to September 2014 with another 5% of the surviving participants. In addition to an updated version of the baseline questionnaire, further data was collected from resurvey participants such as updated blood and urine samples as well as measurements of heel bone density, 12 lead ECG, pulse wave velocity, handgrip strength, ankle-brachial index, and carotid intima

media thickness and plaque measurements. A third resurvey of surviving participants in the CKB is currently underway and will provide even more detailed follow-up data.

All CKB participants have been continuously followed for cause-specific mortality and morbidity, as well as for any hospital admissions detected through linkages with death registries, established registries of major diseases, and health insurance records (covering >97% of participants). In particular, every hospital admission and death is recorded in CKB with a timestamp and one or more disease endpoint codes that indicate the relevant morbidities and/or cause of death.

Ethical approval for CKB was obtained from the ethics committees of the University of Oxford and the Chinese Center for Disease Control and Prevention [79]. Access to CKB data is available at <https://www.ckbiobank.org/site/Data+Access>. For this thesis, we used CKB data releases 17.00 (for Chapters 4-7) and 17.02 (for Chapter 8), which have been fully anonymised and updated with all available follow-up information until January 1, 2018 and include 512,726 of the original 512,891 participants.

3.2 Validation and Adjudication of Stroke End-points

All fatal and non-fatal stroke cases in CKB were coded using ICD-10 by trained medical staff, who were blinded to other personal information, with additional checking and review conducted centrally by trained medical staff. IS (ICD-10 code I63), including lacunar infarction and non-lacunar infarction, was defined as a focal neurological dysfunction lasting for more than 24 hours with or without neuroimaging evidence of a cerebral infarct. HS was defined to include ICH (ICD-10 code I61) and SAH (ICD-10 code I60). ICH was defined as neurological dysfunction caused by haemorrhage into the brain parenchyma or the ventricular system, excluding those induced by injury, with or without neuroimaging evidence of brain haemorrhage. SAH was defined as neurological dysfunction caused by haemorrhage into the subarachnoid space, excluding those induced by injury, with or without neuroimaging evidence of such haemorrhage [80]. Brain imaging (computerised

tomography [CT] and/or magnetic resonance imaging [MRI]) was available for 92% of stroke cases. In the absence of neuroimaging, the diagnosis of stroke types was based on clinical criteria. IS was identified by clinical phenotypes comprising symptoms including unilateral weakness, dysphasia, homonymous hemianopia, monocular or binocular visual loss, diplopia, visual field deficits, visuospatial disorder, brainstem or cerebellar signs (e.g., nystagmus or ataxia), and unclear patterns of neurological deficit. ICH was identified by clinical phenotypes comprising non-focal neurological symptoms including coma, consciousness problems, and neck stiffness. Differential diagnosis of IS and ICH using clinical phenotypes is not as reliable as neuroimaging evidence and may be a limitation on the adjudication accuracy of the 8% of stroke cases without CT or MRI.

All reported cases of first incident stroke with accessible hospital records (n=45,859) underwent additional validation and clinical adjudication, involving attempted retrieval and review of original medical records and brain imaging reports by clinical specialists in China using a bespoke web-based platform (Turnbull *et al.*; manuscript under review). Verified and adjudication-confirmed primary IS cases were further classified into causative subtypes (e.g., CE, LAA, SAO) using CCS (Turnbull *et al.*; manuscript in preparation). A breakdown of the pathological types and subtypes of these strokes are presented in Figure 3.2. Hospital records were retrieved for 85% of the stroke cases, among which the reporting accuracy was found to be 91%. Of the verified primary stroke cases (n=29,952), the diagnostic accuracy was 81%. The high values of reporting and diagnostic accuracy highlight the reliability and high quality of stroke endpoint data in CKB, making it a unique opportunity for studying this disease in China.

3.3 Selection of CKB Data

In order to study stroke risk prediction, we selected a subset of the data available in CKB based on a few key criteria. First, while most of the models developed in this thesis focus on prevention of first-ever stroke cases in alive and previously stroke-free individuals, the investigations described in Chapter 8 also analyse recurrent stroke

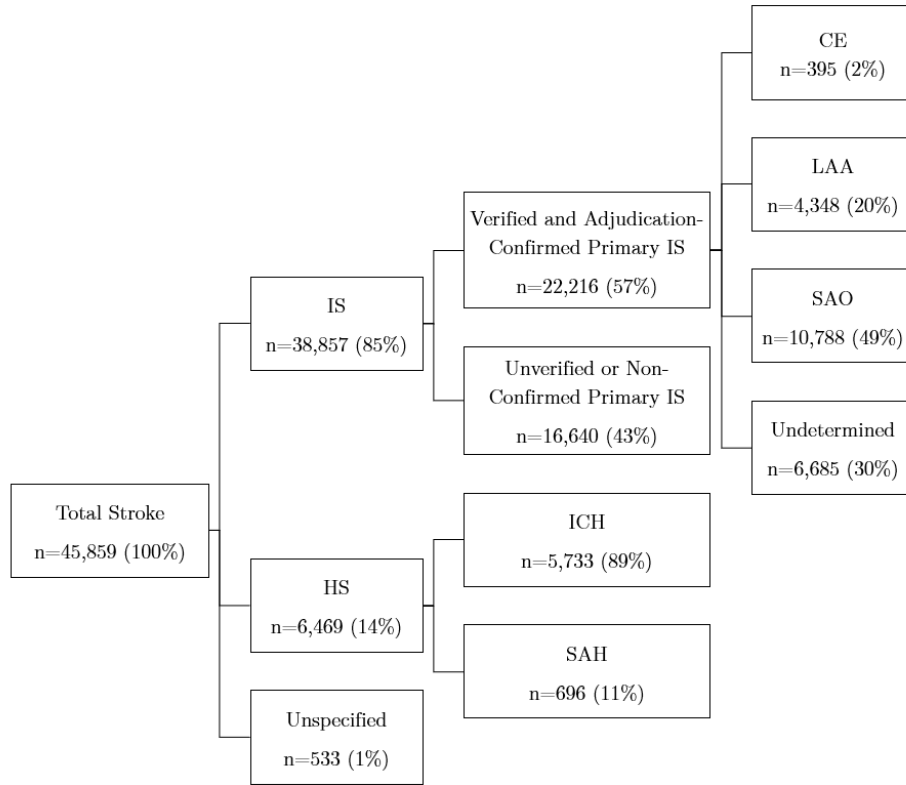


Figure 3.2: Breakdown of stroke pathologies in CKB for 45,859 reported cases of first incident stroke with accessible hospital records. Reported numbers for CE, LAA, and SAO are aggregates of “evident,” “probable,” and “possible” designations assigned by CCS.

rates after various IS subtypes. Consequently, data was requested on strokes or TIA reported prior to the baseline survey as well as the occurrence and timing of all recorded stroke cases or deaths, including separate disease endpoint labels for IS, HS, ICH, SAH, and CCS-classified IS subtypes. Second, to maximise the sample size of individuals with complete data, we focused on exposures for which data was collected at initial recruitment and excluded additional measurements that were only introduced later at the first or second resurvey. However, for the selected exposures, measurements from all resurveys were requested to explore the potential benefit of using sequential risk factor measurements for improved risk prediction (Chapter 6). Third, because our goal was to develop risk models that can be widely implemented in clinical settings upon initial participant screening, we chose to exclude from our models any features that would require further testing such as cardiac test findings, genetic information, or brain imaging data. Finally,

to limit our feature space to only inputs with high-quality data and a reasonably suspected relationship with stroke, a public health expert well-versed in the CKB dataset was consulted to assist in selecting our exposures of interest.

Using these selection criteria, we narrowed and pre-processed our selected data to include a total of 144 original and derived risk factor indicators. A full listing of risk factor indicators used throughout this thesis are provided in the Appendix (Table A.1). The included risk factor indicators used as model inputs throughout this thesis included a range of well-established risk factors used in the Framingham stroke risk models such as age, sex, blood pressure, antihypertensive therapy, diabetes, smoking, and prior CVD. Risk factor indicators were also derived to match the engineered features included in the FSRP (e.g., “SBP per 10 mmHg, if no HRx” and “DM, if Age<65”). One limitation is that some established risk factors for stroke such as AF, left ventricular hypertrophy, and cholesterol levels were not available in CKB. While exclusion of these risk factors may affect the performance of the models developed in this thesis, it is also true that models that do not rely on inputs from cardiac tests or blood tests can be more easily implemented for population-wide risk screening purposes, especially in low resource settings. Meanwhile, the CKB risk factor indicators included many easily-obtainable variables that were not included in the FSRP such as geographic region, socioeconomic indicators, alcohol and smoking details, diet, self- and comparative- health ratings, personal and family medical history, activity level metrics, physical measurements, and heart rate. Additional derived risk factor indicators were included to indicate whether or not each individual was missing reported medical histories for their mother, father, or siblings.

For the investigations described in Chapter 8, we further requested the adjudication data used to classify IS subtypes in CKB using CCS. A full listing of the requested adjudication variables are provided in the Appendix (Table A.2), and include findings from brain imaging, vascular tests, and cardiac tests, as well as the presence of clinical lacunar syndromes.

3.4 Descriptive Statistics of the China Kadoorie Biobank

Of the 512,726 CKB participants, 503,842 individuals (205,293 men, 298,549 women) had no prior history of stroke or TIA at the date of recruitment, but by January 1, 2018, 57,802 incident first-ever stroke cases were recorded in these individuals (48,917 IS and 9,808 HS). In this section, we provide distributions of the most well-established stroke risk factors and highlight their associations with stroke in this previously stroke-free population. We then conclude the section with an analysis and discussion of missing values and censoring in CKB.

3.4.1 Distributions of Well-Established Stroke Risk Factors

3.4.1.1 Age

The baseline age distributions for men and women with and without stroke during the follow-up period are shown in Figure 3.3. The mean age of those with a stroke event was greater than for those without stroke in both sexes (61 years vs 52 years in men, 60 years vs 51 years in women). In addition, for both sexes, there was a significant deficit of participants in their mid-40s, corresponding to the Great Chinese Famine of 1959-1961 [79].

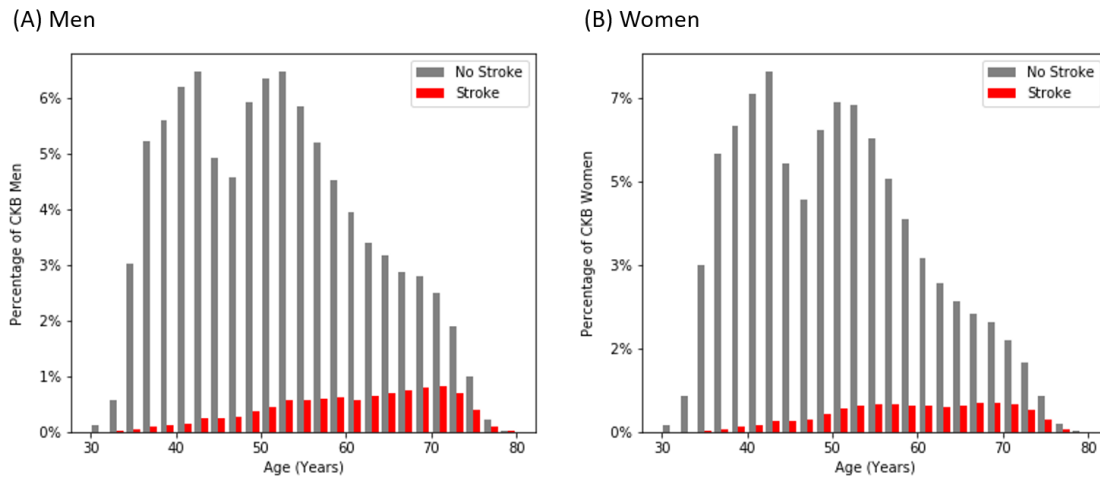


Figure 3.3: Age distributions at baseline survey for (A) men and (B) women with and without stroke during the follow-up period.

3.4.1.2 Blood Pressure and Anti-hypertensive Medication

In CKB, 9.9% of men and 11.4% of women reported use of blood pressure medication at the date of recruitment. Distributions of systolic blood pressure at baseline are shown in Figure 3.4, separated by sex and use of blood pressure drugs. For both untreated and treated individuals, mean systolic blood pressure was higher in those with stroke during the follow-up period than in those without stroke (142 mmHg vs 130 mmHg in untreated men, 138 mmHg vs 126 mmHg in untreated women, 153 mmHg vs 148 mmHg in treated men, 155 mmHg vs 150 mmHg in treated women). However, a much larger proportion of treated individuals experienced a stroke event (21.5% of treated men, 18.3% of treated women) compared to those who did not use blood pressure-lowering medication (8.2% of untreated men, 6.6% of untreated women).

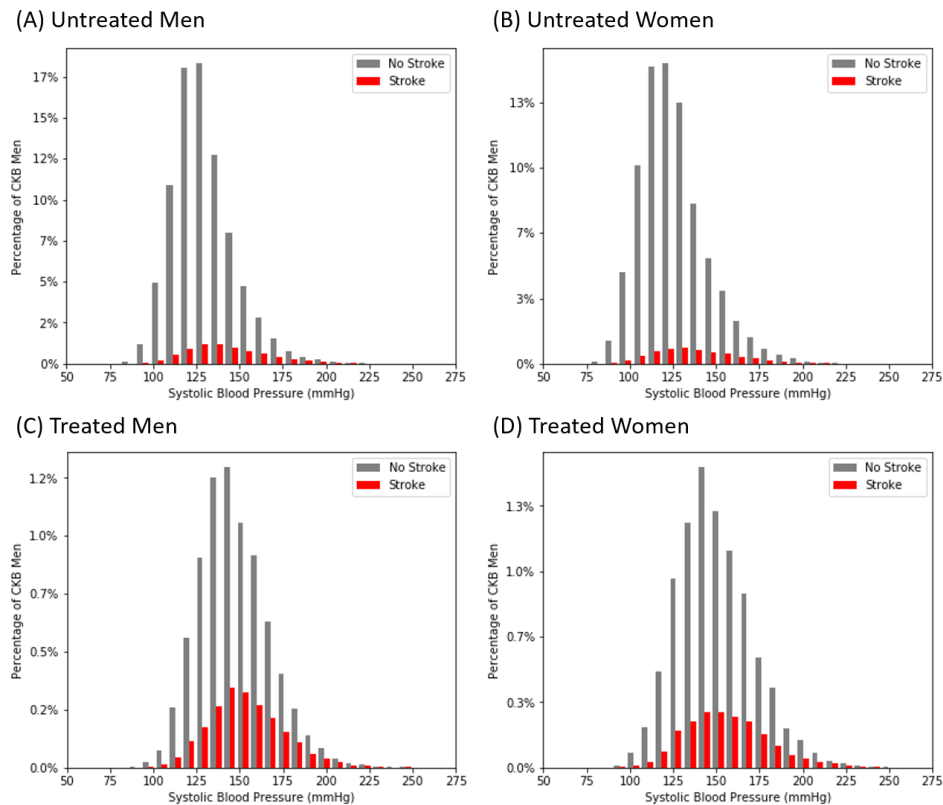


Figure 3.4: Systolic blood pressure (mean of two measurements) for (A) untreated men, (B) untreated women, (C) treated men, and (D) treated women. “Treated”: Reported use of blood pressure drugs. “Untreated”: No reported use of blood pressure drugs.

3.4.1.3 Diabetes

Prevalence of diabetes mellitus (either self-reported or detected) was 5.3% in men and 6.0% in women (Figure 3.5). Among men with diabetes, 20.2% had a stroke, while only 8.9% of men without diabetes had a stroke event. Similarly, 18.7% of women with diabetes had a stroke, while only 7.2% of women without diabetes had a stroke.

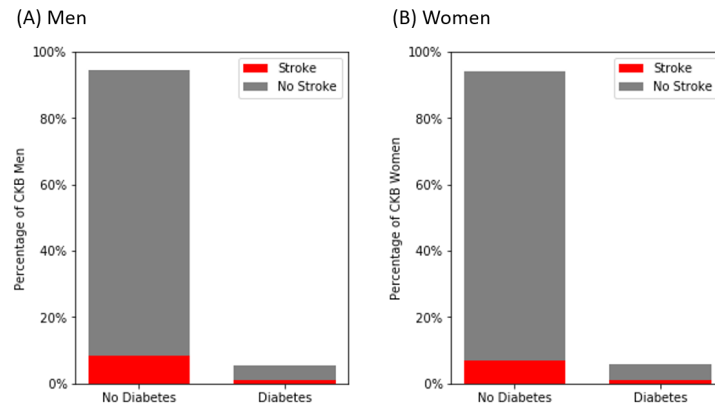


Figure 3.5: Diabetes mellitus at baseline for (A) men and (B) women.

3.4.1.4 Prior CHD

Rates of prior history of CHD were also similar between men and women (2.5% vs 3.0%) (Figure 3.6). Among men who reported prior history of CHD, 24.1% had a stroke, while only 9.2% of men without prior CHD had a stroke event. In women, 24.3% of those with prior history of CHD had a stroke, while the rate was only 7.4% in those without prior CHD.

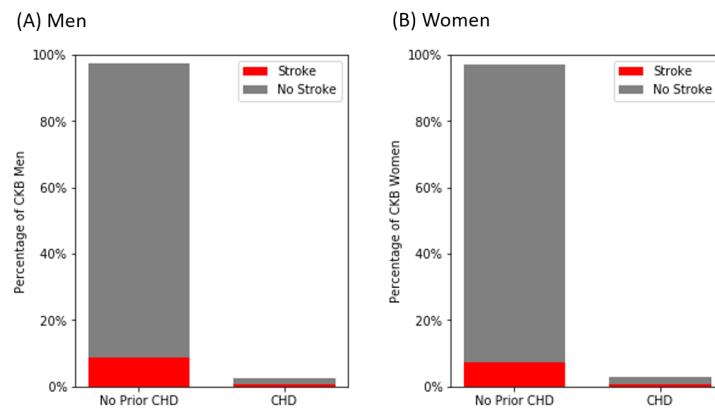


Figure 3.6: Prior CHD at baseline for (A) men and (B) women.

3.4.1.5 Smoking

Rates of current smoking at baseline were 67.7% in men, but only 3.2% in women (Figure 3.7), confirming the well-known differences in smoking patterns between men and women in China [33]. Of those who did smoke, the majority smoked daily. Interestingly, 11.8% of men who were non-smokers at baseline had stroke, which was even greater than the rate of stroke in daily smokers (8.6%). However, in women, this effect was not observed, with 7.8% of non-smokers and 13.5% of daily smokers having a stroke event. The higher stroke rates in men who were non-current smokers may be explained, at least partially, by the inclusion of a substantial number of former smokers in CKB, who stopped smoking by choice or because of illness [81]. Moreover, men who were daily smokers also tended to be younger than non-smokers (mean age: 51 years vs 55 years). This was not true of women, who had a mean age of 59 years among daily smokers and 51 years among non-smokers.

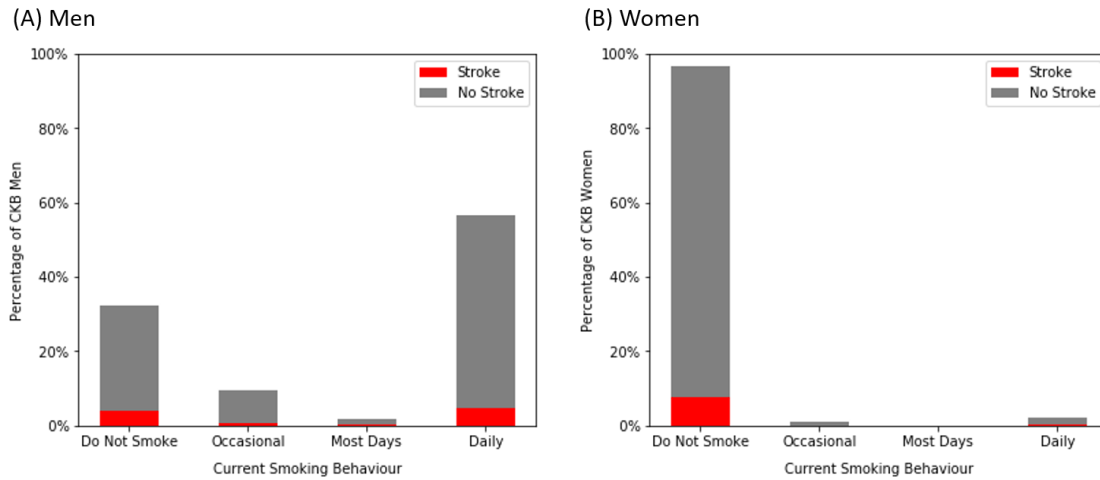


Figure 3.7: Current smoking behaviours at baseline for (A) men and (B) women.

3.4.1.6 Urban and Rural Areas

For both men and women, a higher proportion of participants were recruited from rural areas (57% in men, 56% in women) compared to urban areas (43% in men, 44% in women) (Figure 3.8). However, the proportion of individuals who had a stroke was higher in urban areas (10.9% in men, 9.7% in women) than in rural

areas (8.5% in men, 6.5% in women). This resulted in comparable numbers of strokes being recorded in each setting.

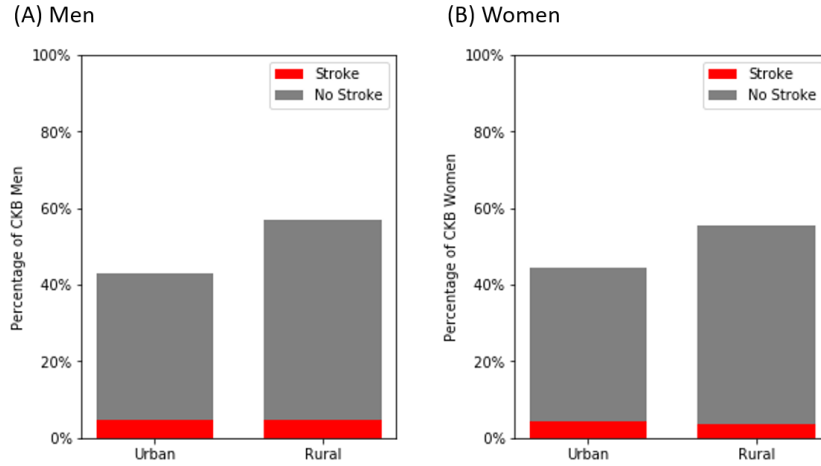


Figure 3.8: Proportion of participants recruited from urban vs rural areas for (A) men and (B) women.

3.4.2 Missing Data and Censoring

In addition to its high-quality stroke data and risk factor information, the CKB is a valuable dataset because of its completeness and limited loss to follow-up (i.e., participants who were right-censored before January 1, 2018 for reasons other than death).

Out of all 144 risk factor indicators used in this thesis (Appendix: Table A.1), only 16 had any missing values. For the 503,842 individuals who had no prior history of stroke or TIA at baseline, 521 individuals had missing values for number of siblings and siblings' medical history (stroke, heart attack, diabetes, and cancer); 817 had missing values for mother's medical history (stroke, heart attack, diabetes, and cancer); 1,124 had missing values for father's medical history (stroke, heart attack, diabetes, and cancer); 2 had missing values for weight and BMI; and 228 had missing values for body fat percentage. In other words, the CKB risk factor indicators used throughout this thesis was almost entirely complete, with only 0.01% of values missing. Since the majority of missing values were related to family medical histories, we added 3 additional binary risk factors indicators

to represent whether or not an individual was missing medical history for their mother, father, or siblings, respectively.

With respect to censoring, a detailed breakdown is provided in Figure 3.9. Overall, censoring rates were very low, with 89.8% of CKB individuals having complete follow-up through January 1, 2018. 9.2% of individuals were censored due to death and only 1.0% of individuals were lost to follow-up. Among those who died, the average time from baseline to death was 6.6 years, and for those lost to follow-up, the mean follow-up time was 8.4 years.

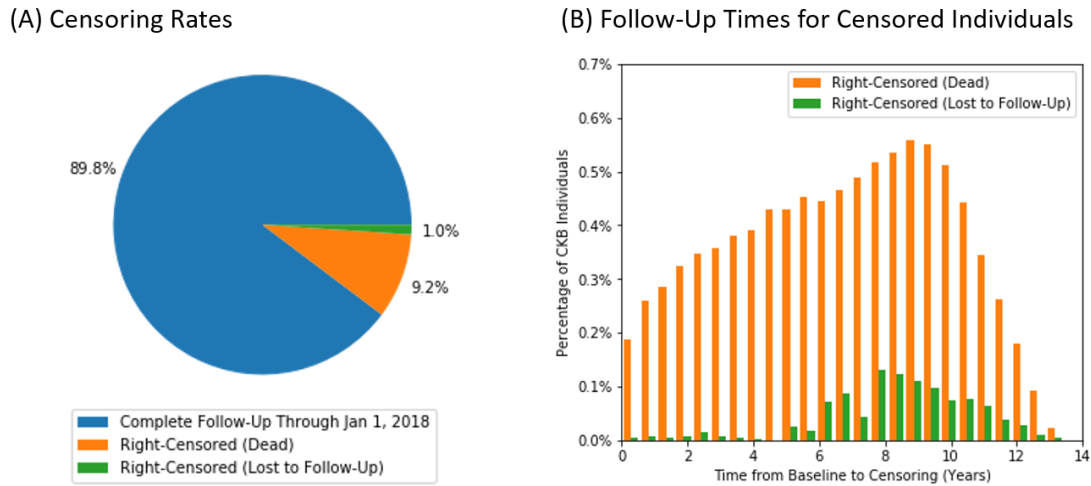


Figure 3.9: Breakdown of censoring in CKB. Subplot (A) shows the censoring rates corresponding to various reasons. Subplot (B) displays the distribution of follow-up times for censored individuals, separated by censoring reason.

3.5 Conclusion

The CKB is one of the largest and most contemporary datasets of its kind in the world and is an invaluable resource for the study of chronic diseases in China. In particular, a combination of (i) high-quality, clinically adjudicated stroke endpoints; (ii) the availability of well-established stroke risk factors and numerous additional variables; and (iii) the impressive completeness of its data makes CKB well-suited for developing state-of-the-art risk prediction models for stroke in the Chinese population. Although CKB was not designed to be representative of the entire Chinese population and lacks data on certain established stroke risk factors (e.g.,

AF, left ventricular hypertrophy, cholesterol, etc.), we hypothesise that stroke risk models developed in CKB will still be equally, if not more applicable to contemporary Chinese adults than existing stroke risk models derived from mostly white cohorts in HICs or outdated Chinese populations. In the remainder of this thesis, we describe the development of such models and share new insights gained about stroke and its various pathologies from this unique prospective cohort study.

4

Traditional Prediction of 9-Year Stroke Risk

4.1 Introduction

As discussed in Chapter 2, traditional models for stroke risk prediction such as the widely-used FSRP [8] have typically been derived using Cox proportional hazards models that have been developed and validated in cohorts from HICs, but have uncertain clinical utility in LMICs such as China [30, 70]. Meanwhile, previous studies that have estimated absolute risk equations for stroke in ethnically Chinese individuals were constrained by insufficient numbers of stroke cases, involvement of single rather than multiple regions, lack of reliable information on stroke types, and lack of contemporary evidence [10, 55, 71].

In this chapter, we first provide an overview of survival analysis and Cox regression, and then describe investigations using data from a large prospective study of 0.5M adults recruited into the CKB in 2004-2008 to compare the performance of the established FSRP and novel Cox risk equations for predicting the absolute risks of total stroke and stroke pathological types in Chinese adults. In particular, we compared the predictive value of (i) the 2017 FSRP; (ii) a recalibrated FSRP; (iii) a recalibrated and refitted FSRP; (iv) a recalibrated and refitted FSRP after stratifying by geographic area; and (v) area-stratified, recalibrated, and refitted

Cox models with additional risk factors for prediction of total stroke, IS, and HS in men and women living in China.

4.1.1 Survival Analysis and Cox Regression

Survival analysis is a branch of statistics concerned with analysing the duration of time until an event of interest occurs (i.e., the “time-to-event”). In the case of strokes, if T is a random variable representing time to an incident stroke case, the distribution of T may be described by its probability density function, $f(t)$, and cumulative distribution function, $F(t)$, as

$$F(t) = \int_0^t f(u) du \quad (4.1)$$

The survival function $S(t)$, which describes the probability that an individual will survive stroke-free beyond time t , is complementary to $F(t)$ as

$$S(t) = P\{T \geq t\} = \int_t^\infty f(u) du \quad (4.2)$$

If $F(t)$ is continuous, then $S(t)$ is equal to $1 - F(t)$. The hazard function, $h(t)$, captures the instantaneous failure rate, or the probability that an individual will have a stroke event in the next instant given that the person survives to t .

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} \quad (4.3)$$

The cumulative hazard function, $\Lambda(t)$, is then given by the accumulation of hazard over time.

$$\Lambda(t) = \int_0^t h(u) du = -\log(S(t)) \quad (4.4)$$

Hence,

$$h(t) = -\frac{d}{dt} \log(S(t)) \quad (4.5)$$

$$S(t) = \exp(-\Lambda(t)) \quad (4.6)$$

The Cox proportional hazards model, or Cox model, is a regression model commonly used for survival analysis. As its name suggests, the Cox model assumes

proportional hazards, meaning the ratio of hazards between the exposed and unexposed groups is assumed to be constant over time even though the underlying hazards may change. The proportional hazards assumption can be tested in multiple ways including observing whether the survival curves intersect or by performing a statistical test based on the Schoenfeld residuals [82].

Cox regression can be formulated as follows. If $\mathbf{Z}_i = \{z_1, z_2, \dots, z_k\}$ is a vector representing k stroke risk factors (such as age, systolic blood pressure, etc.) for the i^{th} individual and $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k\}$ denotes the coefficients, a hazard function of the risk factors for the i^{th} individual can be derived as:

$$h_i(t) = h_0(t)\exp(\hat{\boldsymbol{\beta}}^T \mathbf{Z}_i) \quad (4.7)$$

where $h_0(t)$ is the baseline hazard. Thus, if we assume a single covariate (i.e. risk factor) with a value of 1 in the exposed group and 0 in the unexposed group, then the hazard ratio of the two groups, $\frac{h_{exposed}(t)}{h_{unexposed}(t)}$, is proportional to $\exp(\hat{\boldsymbol{\beta}})$ and does not depend on t . Hence the Cox model is a linear model of the log of the hazard ratio, and allows for hazard ratios to be calculated without specifying a baseline hazard function, making it a semi-parametric model [83].

Although the coefficients, $\hat{\boldsymbol{\beta}}$, of the Cox model can be estimated without a baseline hazard function, absolute risk prediction requires knowledge of the baseline hazard. For this purpose, the non-parametric Breslow estimator [84, 85] for the baseline cumulative hazard function can be calculated as

$$\hat{\Lambda}_0(t) = \sum_{x < t} \left[\frac{dN(x)}{\sum_{i=1}^n Y_i(x)\exp(\hat{\boldsymbol{\beta}}^T \mathbf{Z}_i)} \right] \quad (4.8)$$

where $dN(x)$ is the marginal number of strokes that occur within timestep x , n is the total number of individuals, and $Y_i(x)$ is an indicator variable taking a value of 1 if the i^{th} individual is at risk at timestep x , or 0 otherwise.

If we define $\mathbf{z}^*$ as a k -dimensional vector representing the mean value for each of the k risk factors averaged across all individuals, then a survival function for an “average” individual having mean values for each risk factor can be derived as

$$\hat{S}(t|\mathbf{z}^*) = \exp \left[-\hat{\Lambda}_0(t)\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z}^*) \right] \quad (4.9)$$

Finally, the probability of stroke for the i th individual can be calculated as

$$p_i(\text{stroke}) = 1 - \hat{S}(t|\mathbf{z}^*)^{\exp(\hat{\beta}^T \mathbf{z}_i - \hat{\beta}^T \mathbf{z}^*)} \quad (4.10)$$

4.2 Methods

4.2.1 Inclusion Criteria and Data Preprocessing

Details of the design and methods used in the China Kadoorie Biobank (CKB) are reported in detail in Chapter 3. The analyses presented in this chapter were restricted to CKB individuals with no prior history of stroke or TIA at the date of recruitment (205,293 men, 298,549 women). The participants were followed-up to detect stroke and death until January 1, 2018, and all incident cases of first stroke that were recorded for up to 9 years after the baseline survey were included (19,587 strokes in men; 23,647 strokes in women).

The included CKB individuals were randomly divided into a training set (85%) and test set (15%), stratified by occurrence of total stroke, IS, and HS within 9 years of the baseline survey. Missing values in both the training set and test set (see Section 3.4.2) were imputed using the sex-specific means of the non-missing values in the training set. A listing of all variables in the dataset are included in the Appendix (Table A.1).

4.2.2 Model Development

The FSRP provides sex-specific models that predict an individual’s risk of total stroke within a given time t , based on a Cox model that includes age, current smoking, history of CHD, AF, diabetes, systolic blood pressure, and use of antihypertensive treatment [8]. Therefore, for consistency, all analyses in this chapter were performed separately in men and women. No major violations of the proportional hazards assumption were identified (Appendix B.1).

The FSRP (excluding AF, which was not recorded in CKB) was applied in the test set to predict each individual’s risk of total stroke within 9 years of the baseline survey. A recalibrated model (“+ Recalibration”) was subsequently developed,

using the Breslow estimator [84, 85] to derive a baseline survival function that adjusted for the mean values of risk factors in CKB, while retaining the FSRP hazard ratios (HRs). A recalibrated and refitted model (“+ Refitting”) was then constructed using Cox regression to derive new HRs for the FSRP risk factors in CKB. For recalibration and refitting, model parameters were derived from the training set, and all models were evaluated using the test set.

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

After estimating separate area-specific baselines, we developed an expanded model (“+ Additional risk factors”) using the 133 additional risk indicators recorded at baseline in CKB (Appendix: Table A.1), including socio-demographic factors, diet, alcohol consumption, personal and family medical history, physical activity, and physical measurements. A subset of risk factors were selected using 10-fold cross-validated, least absolute shrinkage and selection operator (LASSO) regularisation (a technique that penalises the inclusion of additional risk factors to prevent overfitting) within the training set [86, 87], and the selected risk factors were used to fit an area-stratified Cox model using the complete training set. Evaluation of the fitted model was performed using the test set.

A visual summary of the complete model development process is shown in Figure 4.1. Since the associations of individual risk factors with stroke pathological types are known to differ [19, 20, 80, 88, 89], we also hypothesised that developing separate risk equations for IS and HS could further improve predictive performance compared to a single model for total stroke. Hence, we repeated the above procedures separately for IS and for HS.

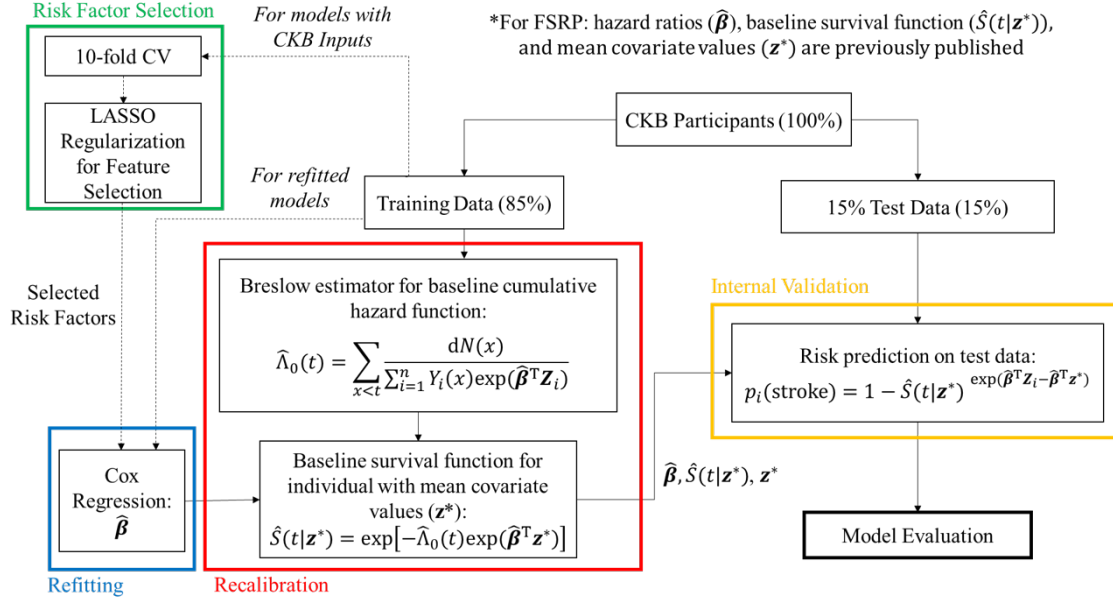


Figure 4.1: Flowchart of the Cox model development process.

4.2.3 Statistical Analysis

To compare predictive performance across models, each model was assessed for risk discrimination and calibration using the test set. Risk discrimination refers to the ability to correctly discriminate between individuals with and without stroke [90], and was evaluated using the area under the receiver operating characteristic curve (AUC). Calibration refers to the similarity between observed and predicted absolute risks [90]. It was evaluated using calibration plots and the Greenwood-Nam-D'Agostino chi-squared test statistic (χ^2), comparing observed incidences (calculated as one minus Kaplan-Meier survival probability) and predicted risks by deciles of predicted risk, and with lower χ^2 values indicating better model calibration [91, 92]. The 95% confidence intervals were constructed for AUCs using 1000 bootstrapped samples from the test set. For models with area-specific baselines, AUCs were evaluated for both the overall study population and separately within each CKB area. Sensitivity analyses were performed using the FSRP [8] and two older Framingham stroke risk equations [50, 51], and for comparability with the 55-84 year age range of the Framingham study, were further restricted to 82,679 men and 105,170 women aged ≥ 55 years. Further sensitivity analyses included

the addition of other potential risk factors to the FSRP prior to stratification by study area, and consideration of competing risk of death using the cumulative incidence function and Fine-Gray model [93].

LASSO variable selection and Fine-Gray analyses were performed in R version 3.6.1 using the glmnet package version 3.0-2 [86] and riskRegression package version 2020.12.08 [94], respectively. All other statistical analyses were performed using Python version 3.7.0. Cox proportional hazards models were implemented using the lifelines package version 0.21.1 [95]. AUC analyses were performed using the scikit-learn toolkit version 0.19.2 [96].

4.3 Results

4.3.1 Baseline Characteristics

Table 4.1: Distribution of established risk factors for total stroke and stroke pathological types in men and women in CKB.

Risk factors included in FSRP*	Men				Women			
	No Stroke (n=185,706)	Total Stroke (n=19,587)	IS (n=16,113)	HS (n=4,587)	No Stroke (n=274,902)	Total Stroke (n=23,647)	IS (n=20,197)	HS (n=4,278)
Age, mean, yr	51.8	60.7	60.8	60.8	50.6	59.6	59.7	59.4
Current smoking, %	68.5	59.9	58.7	63.8	3.1	4.8	4.7	5.4
Coronary heart disease, %	2.1	6.4	6.8	4.8	2.5	9.4	10.0	5.6
Age ≥ 65 yrs, %	13.8	39.9	39.9	41.4	10.7	34.0	34.1	34.6
Diabetes at age < 65 yrs, %	3.7	6.5	7.0	4.7	4.0	8.1	8.6	6.3
Diabetes at age ≥ 65 yrs, %	1.1	4.8	5.3	3.8	1.3	6.0	6.3	4.7
BP-lowering treatment, %	8.6	22.3	22.6	22.9	10.1	26.3	26.3	29.4
SBP-untreated, mean, mmHg	130	142	141	148	126	138	137	148
SBP-treated, mean, mmHg	148	153	152	158	150	155	154	162

*AF ws not recorded in CKB

Note: “No Stroke” column includes individuals lost to follow-up before 9 years, and were stroke-free until being censored.

BP: blood pressure; SBP: systolic blood pressure

Among the 503,842 CKB study participants included in the present analyses, the mean (SD) age was 51.9 (10.6) years and 59% were women. During 9 years of follow-up, a total of 43,234 individuals had any first total stroke (irrespective of type); 36,310 had a first IS and 8,865 had a first HS (Table 4.1). The incidence of first total stroke was higher in men than in women (9.5% vs 7.9%) and varied more than 5-fold across the 10 study areas. Compared with those who had no stroke, individuals who had a first stroke were older and more likely to have prior history of CHD, diabetes, or hypertension. Individuals who had HS were more likely to

be current smokers and had higher mean levels of systolic blood pressure than those who had IS. Overall, men and women had similar rates of prior history of CHD (2.5% vs 3.0%), diabetes (5.3% vs 6.0%), and use of blood pressure-lowering medication (9.9% vs 11.4%), but current smoking was much more common in men than in women (67.7% vs 3.2%) (Table 4.1).

4.3.2 Assessment and Update of 2017 Framingham Stroke Risk Profile

The 2017 FSRP had a moderate discrimination for total stroke in CKB (AUC [95%CI]: 0.78 [0.77-0.79] in men, 0.77 [0.76-0.78] in women) (Table 4.2). However, calibration was very poor, and the FSRP substantially underestimated the absolute risk of total stroke (χ^2 : 1,825 in men, 3,053 in women) (Table 4.2; Figure 4.2).

Recalibration did not alter the AUCs, but substantially corrected the calibration of the model (χ^2 : 156 in men, 506 in women). Refitting the HRs for the calibrated equations did not materially improve discrimination (AUC: 0.79 [0.78-0.80] in men, 0.78 [0.77-0.78] in women), but improved calibration (χ^2 : 51 in men, 148 in women).

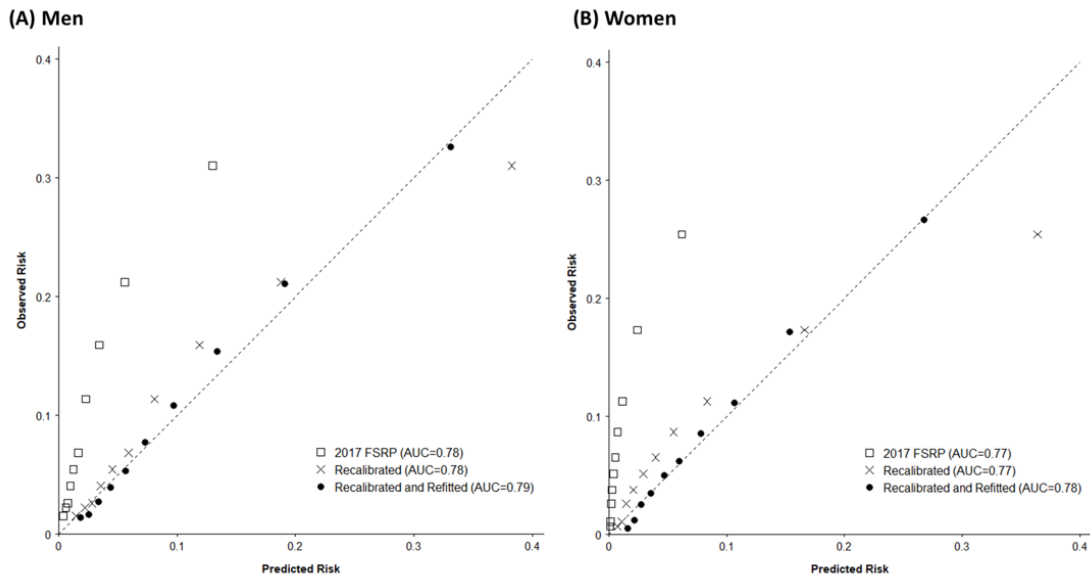


Figure 4.2: Calibration of the 2017 Framingham Stroke Risk Profile (FSRP), and recalibrated and refitted models from China Kadoorie Biobank, for total stroke in men and women. The dashed line in each subplot represents the line of equality between observed risk and predicted risk. Models with better calibration have points lying closer to the line of equality.

Table 4.2: Comparison of performance of different models for prediction of total stroke and stroke pathological types in men and women in CKB.

	Men		Women	
	Discrimination AUC [95% CI]	Calibration χ^2	Discrimination AUC [95% CI]	Calibration χ^2
Total Stroke				
2017 FSRP	0.78 [0.77-0.79]	1,825	0.77 [0.76-0.78]	3,053
+ Recalibration	0.78 [0.77-0.79]	156	0.77 [0.76-0.78]	506
+ Refitting	0.79 [0.79-0.80]	51	0.78 [0.77-0.78]	148
+ Area stratification	0.82 [0.82-0.83]	124	0.82 [0.82-0.83]	178
+ Additional risk factors	0.83 [0.82-0.84]	101	0.83 [0.82-0.84]	177
Ischaemic Stroke (IS)				
2017 FSRP	0.77 [0.76-0.78]	1,200	0.76 [0.76-0.77]	2,406
+ Recalibration	0.77 [0.76-0.78]	118	0.76 [0.76-0.77]	479
+ Refitting	0.78 [0.78-0.79]	21	0.77 [0.76-0.78]	74
+ Area stratification	0.82 [0.81-0.83]	70	0.82 [0.82-0.83]	124
+ Additional risk factors	0.83 [0.82-0.84]	55	0.83 [0.82-0.84]	90
Haemorrhagic Stroke (HS)				
2017 FSRP	0.79 [0.78-0.81]	136	0.78 [0.76-0.80]	70
+ Recalibration	0.79 [0.78-0.81]	58	0.78 [0.76-0.80]	65
+ Refitting	0.80 [0.78-0.81]	23	0.80 [0.78-0.82]	33
+ Area stratification	0.81 [0.80-0.83]	22	0.81 [0.80-0.83]	11
+ Additional risk factors	0.82 [0.81-0.84]	14	0.82 [0.80-0.84]	9

Modifications to models were applied cumulatively. *2017 FSRP*: 2017 Framingham Stroke Risk Profile. *+Recalibration*: Baseline hazard functions re-estimated in CKB. *+Refitting*: Model coefficients from FSRP re-estimated in CKB. *+Area stratification*: Stratification by study area and area-specific estimation of baseline hazard functions. *+Additional risk factors*: Further sociodemographic, health and lifestyle risk factor indicators selected using LASSO regularisation.

4.3.3 Prediction of Total Stroke After Adjusting for Areas in China

Modelling separate area-specific baseline survival functions for total stroke substantially improved risk discrimination among all study participants (AUC: 0.82 [0.82-0.83] in men; 0.82 [0.82-0.83] in women), while maintaining good calibration (χ^2 : 124 in men; 178 in women) (Table 4.2). Discrimination performance within each area is reported in the Appendix (Table B.1). HRs for individual risk factors obtained from these models differed from the FSRP (Figure 4.3) and demonstrated substantially greater consistency between men and women with much greater

precision as reflected by the narrower confidence intervals (as the CKB population was 100-fold larger than the Framingham cohort). A sensitivity analysis including age at which ever-regular smokers started smoking had larger HRs associated with ever-regular smoking in men (1.37) and women (1.17), but showed no material improvement in risk prediction for stroke (Appendix: Figure B.3).

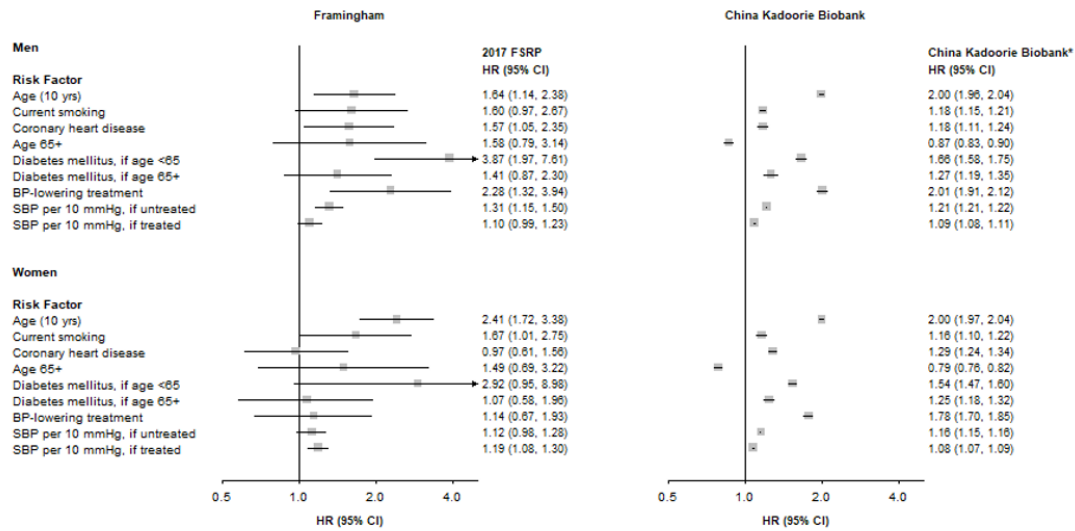


Figure 4.3: Multivariable hazard ratios (HRs) and 95% CI for total stroke in men and women, for Framingham and for CKB. *The 2017 Framingham Stroke Risk Profile coefficients were refitted to CKB in a model including stratification by geographical area.

4.3.4 Expanded Risk Equations with Additional Risk Factors

In addition to controlling for area-specific differences, the addition of other risk indicators recorded in CKB was assessed for risk prediction of total stroke. The expanded models for total stroke, determined using LASSO regularisation for variable selection, included 66 risk factor indicators for men and 70 in women, including measures of diet, personal and family medical history, and socioeconomic status. These models did not yield any further material improvements in either risk discrimination (AUC: 0.83 [0.82-0.84] in men, 0.83 [0.82-0.84] in women) or calibration (χ^2 : 101 in men; 177 in women) (Table 4.2). Discrimination performance within each area is reported in the Appendix (Table B.2).

4.3.5 Risk Equations for Different Stroke Pathological Types

Analysis of separate risk equations for IS and HS demonstrated comparable results from recalibration, refitting, accounting for geographic area, and addition of other risk factors. The best-performing IS model yielded AUCs (95% CI) of 0.83 (0.82-0.84) in men and 0.83 (0.82-0.84) in women with χ^2 values of 55 and 90, respectively. The best-performing HS model yielded AUCs of 0.82 (0.81-0.84) in men and 0.82 (0.80-0.84) in women with χ^2 values of 14 and 9, respectively (Table 4.2).

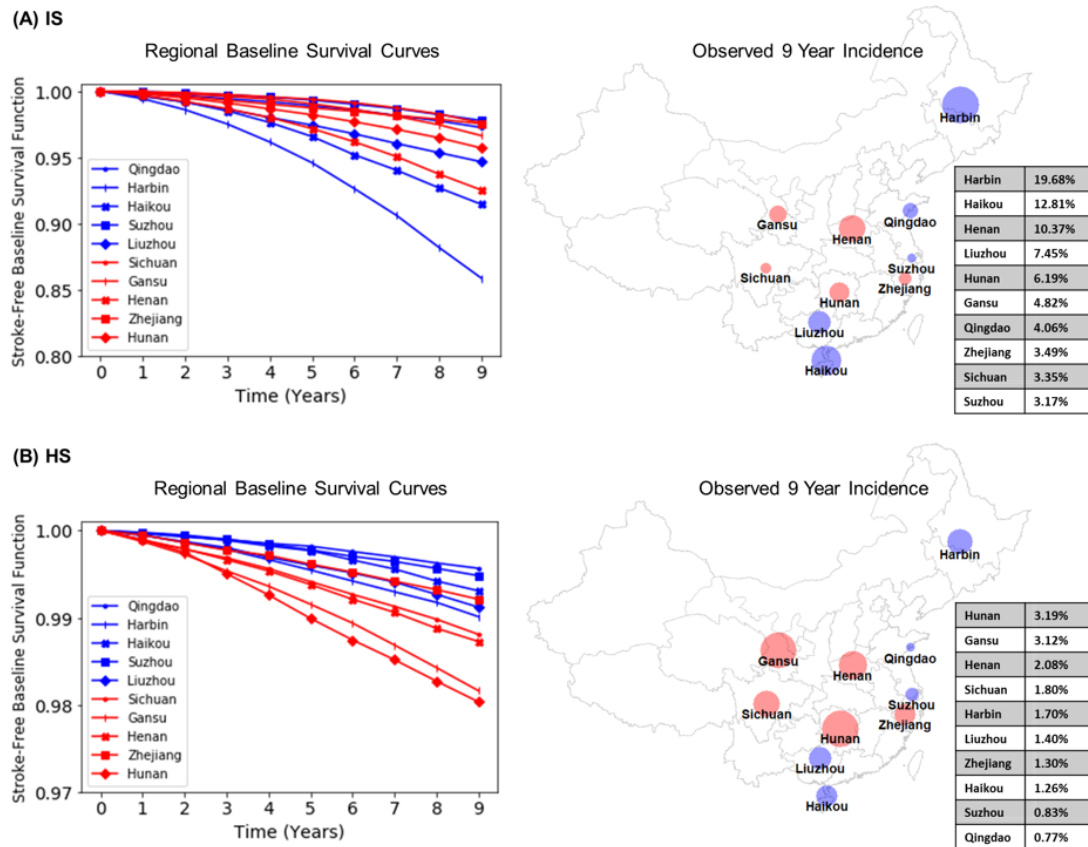


Figure 4.4: Area-specific baseline survival curves and 9 year incidence for (A) IS and (B) HS in CKB. Area-specific baseline survival curves are averaged across sex. Urban study areas are shown in blue while rural areas are shown in red. Note the different scales of the y-axes in subplots (A) and (B). Dot sizes on maps correspond to observed 9 year incidence.

The individual risk equations for IS and HS demonstrated substantial differences between the two stroke pathological types. Overall, the absolute risk of IS was 4 to 5-fold greater than HS, and the ratio of IS to HS risks differed substantially

between areas. Modelling area-specific baseline survival curves for IS and HS demonstrated striking differences in stroke risk between areas, consistent with geographic differences in observed stroke incidence during the 9-year follow-up period (Figures 4.4, 4.5). For example, residents in Harbin had three times the 9-year incidence of IS compared to those in Hunan (19.68% vs 6.19%), but had half the incidence of HS (1.70% vs 3.19%). Furthermore, by training separate models for IS and HS, different HRs were determined for the same risk factors, including CHD (HR for IS/HS; men: 1.18/1.01, women: 1.32/0.97), diabetes ([age <65 years] men: 1.77/1.44, women: 1.60/1.25; [age $\geq$ 65 years] men: 1.34/1.11, women: 1.30/1.02), and blood pressure-related risk factors.

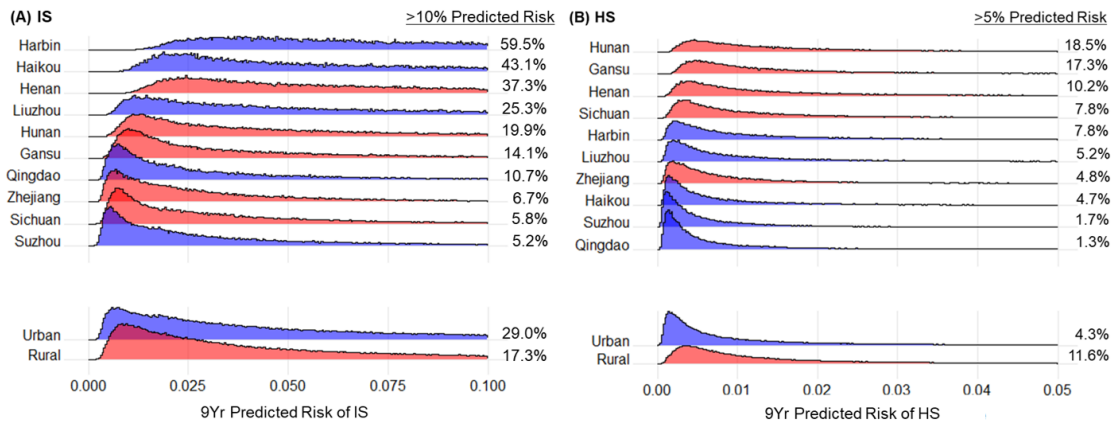


Figure 4.5: Predicted risk distributions for (A) IS and (B) HS by CKB area.

4.3.6 Sensitivity Analyses

Restriction to a subset of participants aged ≥ 55 years yielded comparable results from recalibration, refitting, accounting for area, and including additional risk factors (Appendix: Table B.3). However, due to increased homogeneity among included individuals, the AUCs were lower after excluding the younger adults (age 30-54 years). The age restriction also yielded a greater difference in the AUCs for the best-performing IS and HS models compared to the AUCs of the best-performing total stroke models. Similar results were observed for previous versions of the Framingham stroke risk equations [50, 51] (Appendix: Tables B.4,B.5). Including

additional risk factors prior to area-stratification demonstrated that extra risk factors improved AUCs for total stroke, IS, and HS, with area-stratification contributing to further improved discrimination for total stroke and IS only (Appendix: Table B.6). Stroke incidences adjusted for competing risk of death were very similar to Kaplan-Meier derived estimates (Figure 4.4, Appendix: Figure B.4), and predicted risks from Fine-Gray and Cox models were also similar (Figure 4.6).

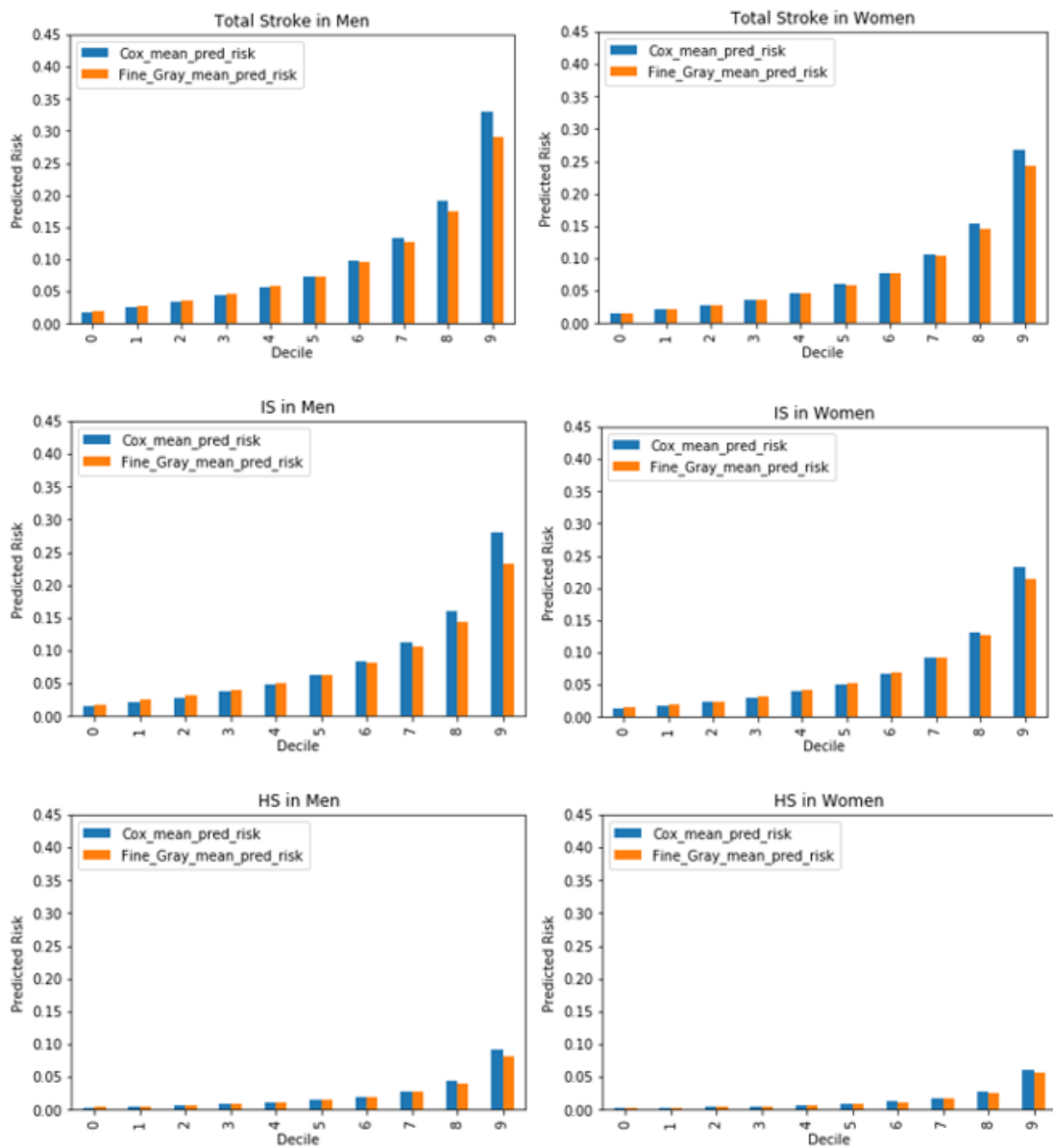


Figure 4.6: Comparison of 9-year risk predictions in test set for Cox model (“+Refitting”) and Fine-Gray model for total stroke, IS, and HS in men and women.

4.4 Discussion

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

A few population-based prospective studies [10, 55, 71] have previously assessed the utilities of FSRP in Chinese adults. In general, they found modest risk discrimination of FSRP for total stroke, but poor prediction of absolute risks of stroke, similar to the present study findings. In the China-PAR study of 106,281 Chinese adults recruited beginning in the early 1990's from 4 cohorts in China with a few thousand stroke events recorded during just over 12 years of follow-up, FSRP yielded modest risk discrimination for total stroke, but greatly underestimated absolute risks of stroke [10]. However, the China-PAR recruited participants as early as the 1990s and no efforts were made in the study to develop new models for improved risk predictions of total stroke and stroke types in more contemporary populations. Likewise, other studies have highlighted the need for recalibration of Framingham-based equations for prediction of cardiovascular disease in LMICs like China [97, 98]. In contrast with previous studies in China, the present study involved a uniquely large number of well-characterised stroke cases (20-fold larger number of stroke cases than the China-PAR study), and developed separate and improved risk prediction equations for total stroke, IS and HS, by taking account of geographical variations in stroke rates within China.

The novel models developed in the present study successfully controlled for area-specific differences that could not be fully explained by the FSRP risk factors alone, and evaluated performance separately for each study area. IS and HS risk distributions were shown to vary significantly by geographic area and in different ways. Some of this variation may be explained by well-known geographic variation in risk factors such as blood pressure [99, 100]. However, some of the variation is unexplained and may be partially due to differences in detection (e.g., from greater use of brain imaging in certain areas). Inclusion of additional risk factors (e.g., socio-demographic factors, alcohol etc.) captured most of the geographic variation in HS risk, but only a fraction for IS risk. Consequently, in studies where controlling for geographic areas is not feasible, the addition of other risk factors that were not included in FSRP could be considered to optimise risk prediction of stroke instead of geographic region.

Current guidelines in both HICs and LMICs advocate the use of blood pressure-lowering medication, lipid-lowering medication, and anti-platelet treatment for cardiovascular disease prevention [101–103]. However, individual subtypes of CVD are heterogeneous in their aetiology, and certain risk factors are known to have heterogeneous effects on individual stroke types [19, 20, 80, 88, 89]. The present study adds to this evidence by highlighting differences in the HRs for risk factors such as CHD, diabetes, and blood pressure-related variables for IS and for HS, and suggests that separately evaluating an individual’s risk for each stroke type may be particularly informative for primary prevention.

Apart from extremely large study population, well-characterised stroke types, and high quality baseline exposure data, a key strength of the present study was the ability to reliably examine and assess the utility of FSRP across different areas simultaneously within the same study. However, the present study also had some limitations. First, atrial fibrillation (AF) was not recorded in CKB, so could not be included in the models. However, other population-based studies of comparable age groups in China [104–106] indicated that the prevalence of AF was substantially lower in China than in Framingham (0.4%–1.7% vs 5.0%), and in

2012, the AF-related stroke prevalence in China was estimated to be 0.13 per 1,000 people [105]. Consequently, although AF is a strong predictor of stroke [104], it is likely to improve risk prediction for only a small portion of individuals in CKB from 2004-2008. As prevalence of AF grows in China [104, 105], it will be increasingly important to incorporate AF into future stroke risk equations. Other limitations of CKB were that recorded stroke events were limited to death from stroke and reported hospitalised strokes (92% having brain imaging to support diagnosis) [80], and CKB is not fully representative of the Chinese population [78, 79]. While the presented risk equations are useful for risk prediction, the HRs for individual risk factors cannot be interpreted causally. Furthermore, it is noted the Cox models presented do not account for competing risk of death and may affect risk estimates, particularly in older individuals. However, with low rates of censoring in CKB (5.4%, with 4.8% of censoring due to death), this effect is small, as evidenced by the very similar 9-year stroke incidences after adjusting for competing risk of death and similar predicted risks between the Cox and Fine-Gray models. Meanwhile, Cox models provide benefits such as direct comparison to the FSRP and cause-specific HRs. Finally, the risk equations outlined in the present study were not designed for immediate implementation in clinical practice, which would require additional validation in independent populations in China and potentially other LMICs.

4.5 Conclusion

The study described in this chapter developed novel risk equations for total stroke, IS, and HS and demonstrated substantial improvements over the widely-used 2017 Framingham Stroke Risk Profile in Chinese adults. Improvements in stroke risk prediction can be attributed to recalibration of baseline survival, refitting hazard ratios, and accounting for geographic differences in stroke incidence in China. The addition of a large number of other risk factors yielded no further material improvements, but may be useful in other studies when area-specific differences are not readily estimated. These techniques can be implemented to improve risk

prediction in any Chinese or similar populations with unique and geographically-diverse risk profiles for stroke. Moreover, separate risk equations for IS and HS could help to identify individuals at high-risk of a particular stroke pathological type and better inform treatment decisions for primary prevention, as described in further detail in Chapter 7. Future work should include the validation and refinement of these equations in independent populations before implementing them for prediction of stroke risk in clinical practice.

5

Machine Learning for Stroke Risk Prediction

5.1 Introduction

Having investigated the performance of traditional Cox models for stroke risk prediction in the China Kadoorie Biobank (Chapter 4), we now turn our attention to the use of machine learning (ML) approaches. ML techniques have been increasingly used in recent years for a variety of healthcare applications, and have demonstrated superior predictive value compared with Cox models for predicting risk of stroke and overall CVD in various populations [62–64, 67, 72]. However, these ML models have still not been widely adopted in clinical practice and, once again, little is known about the utility of such risk scores for prediction of stroke risk in a contemporary Chinese population [66].

In this chapter, we provide an overview of various ML techniques and evaluate their utility for stroke risk prediction in CKB. The objectives of the investigations in this chapter were to (i) compare Cox and ML models for prediction of risk of stroke in China at varying intervals of follow-up (i.e., stroke within 9 years, 0-3 years, 3-6 years, 6-9 years); (ii) identify individuals for whom ML models might be superior to conventional Cox-based approaches for stroke risk prediction;

and (iii) develop and evaluate an ensemble model combining both approaches to identify individuals at high-risk of stroke.

5.1.1 Overview of ML Approaches

5.1.1.1 Regularised Logistic Regression (LR)

Similar to the Cox model (described in Section 4.1.1), logistic regression (LR) is another form of multiple linear regression that assumes (i) the system to be represented is linear, and (ii) the system can be described parametrically using coefficients with scalar values. The aim is to find coefficient values that maximise the accuracy of transforming the input data (e.g., risk factors) to the desired output value (e.g., probability of having a stroke within 9 years) [107].

LR can be used to predict the i^{th} individual's probability of having a stroke, $p_i(\text{stroke})$, as

$$p_i(\text{stroke}) = \frac{1}{1 + e^{-Z}} \quad (5.1)$$

$$Z = w_1x_{i,1} + w_2x_{i,2} + \dots + w_kx_{i,k} \quad (5.2)$$

where w_j is the j^{th} coefficient to be estimated, $x_{i,j}$ is the value of the j^{th} risk factor for individual i , and k is the total number of risk factors used to fit Z . The coefficient vector, $\mathbf{w} = \{w_1, w_2, \dots, w_k\}$, is found by minimising the cross-entropy loss, which can be expressed as

$$\text{Cross-entropy Loss} = -\frac{1}{n} \sum_{i=1}^n [(y_i)\log(p_i(\text{stroke})) + (1 - y_i)\log(1 - p_i(\text{stroke}))] \quad (5.3)$$

where n is the total number of individuals, and y_i is the true target to be predicted for individual i , taking a value of 1 if individual i has a stroke or 0 otherwise.

To encourage model sparsity and prevent overfitting, feature selection can be performed by modifying the loss function (Equation 5.3) to include a regularisation term. Throughout this chapter, we use L1, or LASSO, regularisation, which pushes

coefficients of non-predictive risk factors to 0 by penalising the absolute value of the magnitude of the coefficients. The modified L1-regularised cost function is given by

L1-Regularised Loss =

$$-\frac{1}{n} \sum_{i=1}^n [(y_i) \log(p_i(\text{stroke})) + (1 - y_i) \log(1 - p_i(\text{stroke}))] + \frac{\lambda}{n} \sum_{j=1}^k |w_j| \quad (5.4)$$

where λ is a hyperparameter representing the strength of regularisation and can be tuned using cross-validation.

After minimising the loss function, the derived coefficient vector, $\mathbf{w}$, is then used in the logistic function defined by Equation 5.1, which converts the values of Z to a predicted probability, $0 \leq p_i(\text{stroke}) \leq 1$, for each individual.

5.1.1.2 Support Vector Machines (SVM)

The support vector machine (SVM) is a non-probabilistic classifier that defines a hyperplane that is as wide as possible between the separate classes, otherwise known as a “wide margin” classifier. In this chapter, we use the standard SVM, which is a linear classifier [107].

Given a dataset of the form $(\mathbf{x}_1, y_1) \dots (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i$ is the feature vector of input data for individual i , and y_i is the corresponding class $\{-1, 1\}$, a hyperplane can be specified which separates the classes,

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \quad (5.5)$$

where $\mathbf{w}$ is a coefficient vector normal to the hyperplane, and b is a bias term. The data points closest to the hyperplane from each class are referred to as “support vectors”, and two parallel hyperplanes are specified that run through these points. These parallel hyperplanes are described by:

$$\mathbf{w} \cdot \mathbf{x} - b = 1 \text{ and } \mathbf{w} \cdot \mathbf{x} - b = -1 \quad (5.6)$$

The distance between the two hyperplanes is $\frac{2}{\|\mathbf{w}\|}$ so the best classification is determined by minimising $\|\mathbf{w}\|$, subject to $y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1$ for all individuals i . In some cases, a “soft-margin” may be implemented, using a regularisation

hyperparameter C which penalises misclassified points in proportion to their distance from the correct hyperplane. Thus, a smaller value of C corresponds to a “softer” margin and lower penalisation of misclassified points.

To calibrate the non-probabilistic SVM outputs to probabilities, techniques such as Platt scaling or isotonic regression can be used [108]. In this chapter, we implement isotonic regression, which is a non-parametric technique for fitting a non-decreasing function that maps the non-probabilistic SVM outputs to calibrated probabilities.

5.1.1.3 Decision Trees

Decision trees are non-parametric ML techniques used for classification and regression. Various decision tree algorithms exist, of which we implement “Classification and Regression Trees” (CART) in this chapter. The CART algorithm constructs binary trees, with splits created at specific thresholds on individual features in order to maximise information gain at each node [109].

The data at node d can be represented by Q_d with N_d samples. Each sample represents an individual and can be expressed as $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is a feature vector for individual i and y_i is the corresponding label value (0 for no stroke, 1 for stroke). For each candidate split $\theta = (f, t_d)$ consisting of a feature f and threshold t_d , the data is partitioned into two subsets, $Q_d^{left}(\theta)$ and $Q_d^{right}(\theta)$, where

$$Q_d^{left}(\theta) = \{(\mathbf{x}, \mathbf{y}) | x_{i,f} \leq t_d\} \text{ and } Q_d^{right}(\theta) = Q_d \setminus Q_d^{left}(\theta) \quad (5.7)$$

The quality of the split is evaluated as

$$G(Q_d, \theta) = \frac{N_d^{left}}{N_d} H(Q_d^{left}(\theta)) + \frac{N_d^{right}}{N_d} H(Q_d^{right}(\theta)) \quad (5.8)$$

where $H(Q_d)$ is the Gini impurity:

$$H(Q_d) = \sum_k p_{dk}(1 - p_{dk}) \quad (5.9)$$

and p_{dk} represents the proportion of class k observations (0 for no stroke, 1 for stroke), in node d .

Splits are selected to minimise the Gini impurity and are recursively applied until the maximum allowable depth of the decision tree is reached or until there is

a pre-specified minimum number of samples remaining in node d . The maximum tree depth, minimum number of samples per node, and number of randomly selected features considered at each split are all hyperparameters that can be tuned using cross-validation.

5.1.1.4 Gradient Boosted Trees (GBT)

Gradient Boosted Tree (GBT) models are an extension of decision trees and utilise the concept of “boosting.” Boosting entails the use an ensemble of shallow decision trees, or “weak learners,” to improve predictive performance [110]. The output of a GBT model for an individual i can be expressed as $F_M(\mathbf{x}_i)$, where $F_M(\mathbf{x}_i)$ is the sum of the outputs of M decision tree estimators, h_m :

$$F_M(\mathbf{x}_i) = \sum_m h_m(\mathbf{x}_i) \quad (5.10)$$

The GBT ensemble is built iteratively in a greedy manner:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + h_m(\mathbf{x}) \quad (5.11)$$

where the newly added estimator h_m is fitted in order to minimise a sum of losses L_m , given the previous ensemble F_{m-1} :

$$h_m = \arg \min_h L_m \approx \arg \min_h \sum_{i=1}^n (l(y_i, F_{m-1}(\mathbf{x}_i)) + h(\mathbf{x}_i)g_i) \quad (5.12)$$

$$g_i = \left[\frac{\partial l(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)} \right]_{F=F_{m-1}} \quad (5.13)$$

L_m is minimised when the estimator h_m is fitted to predict the negative gradients (or “pseudo-residuals”) of the samples, $-g_i$. In the case of binary classification, we use the the binary cross-entropy loss function (Equation 5.3) for l and use the log odds ratio for the initial model F_0 . To convert the GBT outputs to meaningful probabilities of an individual belonging to the positive class (i.e., having a stroke), the output $F_M(\mathbf{x}_i)$ is put through the logistic function (Equation 5.1).

In addition to sharing the hyperparameters of decision trees, GBT models should be further tuned using cross-validation to identify an appropriate learning rate (which controls the level of contribution of each new estimator), total number of base estimators, and levels of subsampling used to train each estimator.

5.1.1.5 Multilayer Perceptrons (MLP)

The multilayer perceptron (MLP) is a class of artificial neural networks characterised by fully connected nodes and layers [111]. MLPs comprise an input layer, an output layer, and one or more hidden layers in between. Each node represents operations on data from the previous layer and is fully “feedforward connected,” meaning that the output of each node in a given layer is fed into every node of the subsequent layer.

Input data are typically scaled and/or normalised, and then presented to the MLP at the input layer with each feature representing a node. Each connection from one layer to the next corresponds to a weight, and since the MLP is fully connected, there will be $N_n N_{n+1}$ weights where N_n and N_{n+1} are the number of nodes within layers n and $n + 1$ respectively.

Each node corresponds to an “activation function,” which emulates biological neural networks by creating a binary output at each node, similar to how neurons in a biological network either fire or do not. Examples of activation functions include the sigmoid or logistic function (Equation 5.1), where the output is constrained between $[0,1]$; the rectified linear unit or “ReLU,” $f(x) = \max(0, x)$; and the hyperbolic tan function, $f(x) = \tanh(x)$. The activation function must be differentiable in order to be able to train the network, with computationally efficient differentiation having significant value for large MLPs in which many derivatives must be found.

Given a k -dimensional input feature vector for an individual i , $\mathbf{x}_i = (x_1, \dots, x_k)$, the operation within the q^{th} node of the first hidden layer would be described by:

$$y_q = f \left(\sum_{p=1}^k w_{pq} x_p + b_1 \right) \quad (5.14)$$

where b_1 is a bias term and w_{pq} is the weight from the p th node in the previous layer to the q th node in the subsequent layer.

In the next layer, the operation performed in Equation 5.14 would be repeated using the outputs of the previous layer as current inputs, continuing onward until the final layer. In the case of binary stroke prediction, the final layer consists of a single node with a sigmoid activation function that outputs a probability that the individual i will have a stroke event.

The MLP is trained using backpropagation [111], which uses errors $E = \sum_i (y_i - \hat{y}_i)$ between the target value y_i for individual i and the network's output $\hat{y}_i$ to optimise the value of the weights. Neural networks do not guarantee a convex minimisation problem, meaning that getting stuck at local minima in the optimisation space is a challenge that may need to be overcome [111, 112].

Along with the architecture of the network, the MLP performance can be influenced by the optimisation method used to train the weights. In this chapter, we used mini-batch gradient descent, which can be expressed mathematically as:

$$\mathbf{w}_{n_{iter}+1} = \mathbf{w}_{n_{iter}} - \eta \sum_{i=1}^{n_{batch}} \nabla Q_i(\mathbf{w}_{n_{iter}}) / n_{batch} \quad (5.15)$$

where n_{iter} is the current iteration number, n_{batch} is the number of randomly selected individuals included in the batch, $\mathbf{w}$ is a weight vector, $\nabla Q_i(\mathbf{w}_{n_{iter}})$ is the gradient of the cost function at the i^{th} individual in the batch, and η is the learning rate. In mini-batch gradient descent, the gradient is calculated based on a small batch of randomly selected individuals at each iteration in order to balance computational speed and noise in the estimated gradient.

Important hyperparameters that need to be tuned in MLP models include those specifying the model architecture such as number of hidden layers, number of hidden nodes, types of activation functions, and optional use of batch normalisation. In addition, weight initialisation, learning rate and batch size can all be tuned to control the speed and quality of the learning process. Finally, to prevent overfitting on the training data, a range of techniques are available for tuning including early stopping, dropout, L1 regularisation, and L2 regularisation.

5.1.1.6 Random Survival Forests (RSF)

Moving beyond ML techniques for binary classification, we also considered random survival forests (RSF) [113] as a ML technique for survival analysis (described above in Section 4.1.1). Like the Cox model, RSF can be used effectively on right-censored data and makes time-to-event predictions that can then be converted to a binary prediction of stroke at any particular time point (e.g., within 9 years).

Similar to GBT, RSF utilises an ensemble of tree-based estimators (“survival trees”) to improve predictive performance. However, rather than sequentially training a series of shallow base estimators, RSF utilises a “bagging” approach, in which parallel trees of full depth are developed on various bootstrapped samples of the training data to estimate the cumulative hazard function, with individual estimates being averaged together.

As in CART, each estimator is a binary tree grown through recursive splitting of nodes. At each node, a random selection of N_f candidate features are selected to generate a split that maximises the survival difference between the two daughter nodes, until a pre-specified minimum number of samples is reached. The estimate of the cumulative hazard function, $\Lambda(t)$, at each terminal node d is estimated using the Nelson-Aalen estimator:

$$\hat{\Lambda}_d(t) = \sum_{t_{l,d} \leq t} \frac{strokes_{l,d}}{individuals_{l,d}} \quad (5.16)$$

where $t_{1,d} < t_{2,d} < \dots < t_{N(d),d}$ are $N(d)$ distinct event times, $strokes_{l,d}$ represents the number of strokes at time $t_{l,d}$, and $individuals_{l,d}$ represents the number of individuals at risk at time $t_{l,d}$. The survival estimate, $\hat{S}_d(t)$, for each node d is obtained using:

$$\hat{S}_d(t) = \exp(-\hat{\Lambda}_d(t)) \quad (5.17)$$

Important hyperparameters for training RSF models include the number of survival tree estimators, the number of candidate features (N_f), and the minimum number of samples per node.

5.2 Methods

5.2.1 Inclusion Criteria

Details of the design and methods used in the China Kadoorie Biobank (CKB) are reported in detail in Chapter 3. As in Chapter 4, the analyses presented in this chapter were restricted to CKB individuals with no prior history of stroke or TIA at the date of recruitment (205,293 men, 298,549 women). The participants were

followed-up to detect stroke and death until January 1, 2018, and all incident cases of first stroke that were recorded for up to 9 years after the baseline survey were included (19,587 strokes in men; 23,647 strokes in women).

5.2.2 Model Development and Validation

CKB individuals were randomly assigned to a training set (85%; 174,498 men with 16,649 strokes; 253,766 women with 20,100 strokes), a validation set (12.75%; 26,174 men with 2,467 strokes; 38,065 women with 3,014 strokes), and test set (2.25%; 4,620 men with 471 strokes; 6,718 women with 533 strokes), with all subsequent analyses performed separately by sex (Figure 5.1). Missing values (see Section 3.4.2) were imputed using the sex-specific means of the non-missing values in the training set. A listing of all variables in the dataset are included in the Appendix (Table A.1).

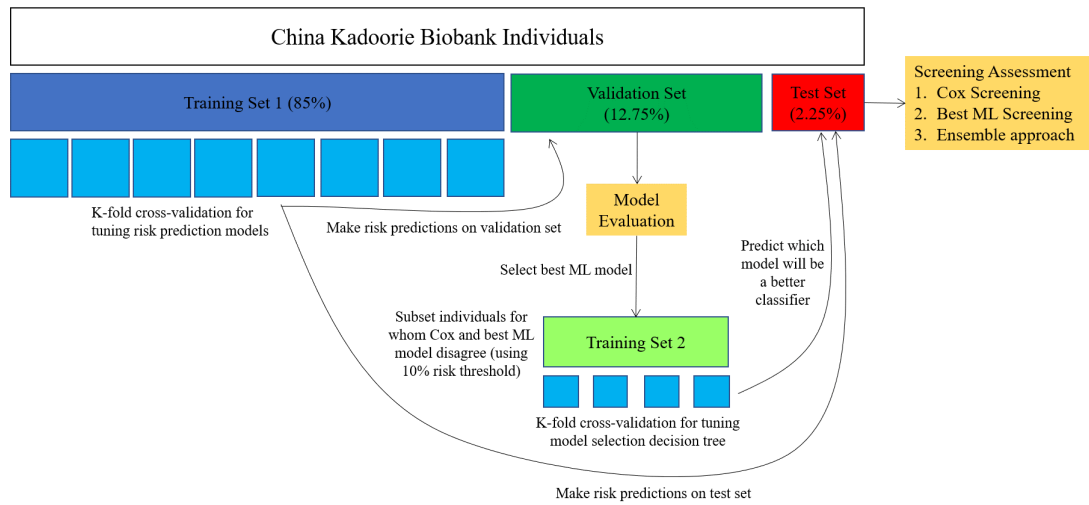


Figure 5.1: Cox and machine learning (ML) model development and validation. All analyses were performed separately for men and women. Included individuals were divided into a training set (85%), validation set (12.75%), and test set (2.25%). Risk prediction models were developed in the training set and assessed in the validation set, with a best ML model selected. The traditional Cox model and best ML model were then used for screening high-risk individuals in the validation set using a 10% predicted risk threshold. A second training set was created from a subset of the validation set wherein the Cox model and best ML model disagreed on risk classification, and a decision tree was trained to predict which model would yield a better risk classification for each individual. Screening approaches including a (i) Cox-only approach, (ii) best ML-only approach, and (iii) and ensemble approach were assessed and compared using the held-out test set.

Cox, random survival forest (RSF), logistic regression (LR), support vector machine (SVM), gradient boosted tree (GBT), and multilayer perceptron (MLP) models were derived in the training set for risk prediction of stroke within 9 years of the baseline survey. To explore differences in performance and major risk factors for short-term and long-term risk prediction, models were also derived for follow-up intervals of 0-3 years, 3-6 years, and 6-9 years after baseline. Features were selected and hyperparameters tuned in each model using k-fold cross validation within the training set (Appendix C.1), and the final models were evaluated in the validation set. All models were benchmarked against the 2017 Framingham Stroke Risk Profile (FSRP) [8], both with and without recalibration and refitting to the CKB cohort. Atrial fibrillation (AF) was not recorded in CKB and was excluded from the FSRP model.

Survival analysis approaches (FSRP, Cox, RSF) differ from binary classifiers (LR, SVM GBT, MLP) in their ability to account for censored individuals and yield time-to-event probabilities for stroke. Consequently, stroke-free individuals who died or were lost to follow-up before 9 years (5.4% of all participants) were included in the training set for development of survival models, but not in the binary classification models. These right-censored individuals were further excluded from the validation set and test set for all models since it was unknown if they could have suffered a stroke within the time interval of interest. Furthermore, while a single Cox and RSF model could be used for risk prediction at all time scales, separate binary classification models were derived for each prediction task.

After comparing model discrimination and calibration, the predictions of the Cox model and best-performing ML model were used to identify individuals at high-risk of stroke (defined as having >10% predicted risk of stroke in 9 years) in the validation set. Agreement between the Cox model and best-performing ML model was assessed qualitatively using t-Distributed Stochastic Neighbor Embedding (t-SNE), an unsupervised, non-linear technique for visualising high-dimensional data. A second training set was generated by restricting the validation set to the individuals for whom the Cox model and best ML model disagreed, and a decision

tree was derived from this training set (Appendix C.1) to predict which model would yield a better classification (i.e., “high-risk” if the individual had a stroke or “not high-risk” if the individual did not have a stroke) based on individual-level characteristics. Feature importance for the decision tree was analysed using the Gini importance metric. Screening approaches to identify individuals at high-risk of stroke were then compared in the test set, including a Cox-only approach, best ML-only approach, and an ensemble approach that used the trained decision tree to select a model in cases of disagreement between Cox and the best ML model.

5.2.3 Statistical Analysis

The Cox and ML models were assessed for risk discrimination and calibration performance in the validation set. Risk discrimination refers to the ability to correctly discriminate between individuals with and without stroke [90], and was evaluated using the area under the receiver operating characteristic curve (AUROC), with higher AUROCs indicating better risk discrimination. Calibration refers to the similarity between observed and predicted numbers of stroke events for each predicted risk decile [90], and was evaluated using chi-squared test statistics (χ^2) from the Hosmer-Lemeshow test (for binary classification models) and Nam-D’Agostino test (for survival models), with lower χ^2 values indicating better calibration [91, 92]. The 95% confidence intervals were estimated for AUROC and χ^2 values using 1000 bootstrapped samples from the validation set.

Risk screening approaches were evaluated in the test set, using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Agreement between approaches was assessed using Cohen’s kappa (κ). For the ensemble screening approach, the decision tree to select between Cox and the best ML model was evaluated for accuracy as well as discriminatory performance using AUROC in the test set.

Statistical analyses were performed using Python version 3.7.0 and R version 3.6.1. Cox models were implemented using the lifelines package version 0.21.1 [95]

with LASSO variable selection performed in R using the `glmnet` package version 3.0-2 [86]. RSF models were implemented using the `ranger` package version 0.12.1 [114]. LR, SVM, and GBT models were implemented using `scikit-learn` toolkit version 0.19.2 [96], and MLP models were implemented using `keras` version 2.3.1 [115]. tSNE visualisations were implemented using the `Rtsne` package version 0.15 [116].

5.3 Results

5.3.1 Baseline Characteristics

The baseline characteristics of included CKB study participants have been previously described in Section 4.3.1. As a brief overview, the mean (SD) age was 51.9 (10.6) years and 59% were women (Table 4.1). During 9 years of follow-up, a total of 43,234 individuals had a first stroke, and the incidence of stroke was higher in men than in women (9.5% vs 7.9%) and varied by more than 5-fold between the 10 study areas. A distribution of strokes over time from baseline survey is shown in Figure 5.2.

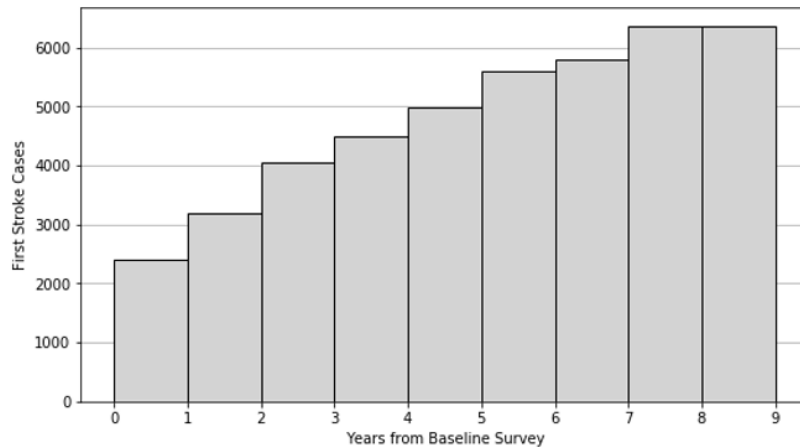


Figure 5.2: Distribution of strokes over time from baseline survey in 503,842 adults with no prior history of stroke in the China Kadoorie Biobank.

Compared with those who had no stroke, individuals who had a first stroke were older and more likely to have prior history of CHD, diabetes, or hypertension (Table 4.1). Overall, men and women had similar proportions with prior history of CHD (2.5% vs 3.0%), diabetes (5.3% vs 6.0%), and use of blood pressure-lowering

medication (9.9% vs 11.4%), but the prevalence of current smoking was much higher in men than in women (67.7% vs 3.2%).

5.3.2 Comparisons of Cox versus ML models to predict risk of stroke

Table 5.1: Discrimination and calibration performance for prediction of 9-year risk of stroke. Comparisons included the 2017 Framingham Stroke Risk Profile (FSRP), a recalibrated and refitted FSRP, Cox, random survival forest (RSF), logistic regression (LR), support vector machine (SVM), gradient boosted tree (GBT), and multilayer perceptron (MLP) models.

Model Type	Men		Women	
	Discrimination	Calibration	Discrimination	Calibration
	AUROC χ^2 [95% CI]	χ^2 [95% CI]	AUROC χ^2 [95% CI]	χ^2 [95% CI]
FSRP	0.781 [0.772-0.790]	5,541 [4,996-6,107]	0.772 [0.764-0.780]	19,402 [17,784-21,019]
Recalibrated and Refitted FSRP	0.824 [0.816-0.831]	138 [96-185]	0.825 [0.819-0.833]	140 [97-186]
Cox	0.829 [0.822-0.837]	122 [83-166]	0.831 [0.824-0.838]	129 [89-172]
RSF	0.826 [0.818-0.834]	61 [36-90]	0.832 [0.824-0.839]	62 [36-95]
LR	0.831 [0.823-0.838]	56 [31-86]	0.832 [0.825-0.838]	57 [34-85]
SVM	0.830 [0.823-0.838]	712 [582-852]	0.831 [0.824-0.838]	24 [11-41]
GBT	0.833 [0.825-0.840]	44 [24-67]	0.836 [0.829-0.843]	47 [30-69]
MLP	0.831 [0.824-0.839]	515 [410-627]	0.833 [0.826-0.841]	19 [8-35]

The Cox model and ML models all outperformed the 2017 Framingham Stroke Risk Profile (FSRP) with and without recalibration and refitting, and achieved similar discrimination for 9-year risk of stroke, with GBT yielding marginally higher AUROCs than other models for both men and women (Table 5.1, Appendix: Figure C.3). However, calibration performance varied substantially between models. RSF, LR, and GBT all yielded lower χ^2 values than Cox models in both men and women, with GBT showing significant improvements in calibration performance (Table 5.1, Appendix: Figure C.4). MLP and SVM (after isotonic regression) were observed to yield good calibration for women but were poorly calibrated for men. Nevertheless, all models were significantly better calibrated than the original FSRP, which severely

underestimated stroke risk in the CKB cohort (Figures 5.3, Appendix: Figure C.4). Calibration plots (Figure 5.3) indicated that calibration for all models was better in women than in men, with models underestimating risk of stroke in men compared with women. Furthermore, deviation from perfect calibration was more extreme in those at highest risk of stroke (Figure 5.3) and in older individuals (Appendix: Figure C.5). Due to its high AUROC and improvements to calibration over the traditional Cox model, GBT was identified as the best-performing ML model and was selected for further risk screening analyses.

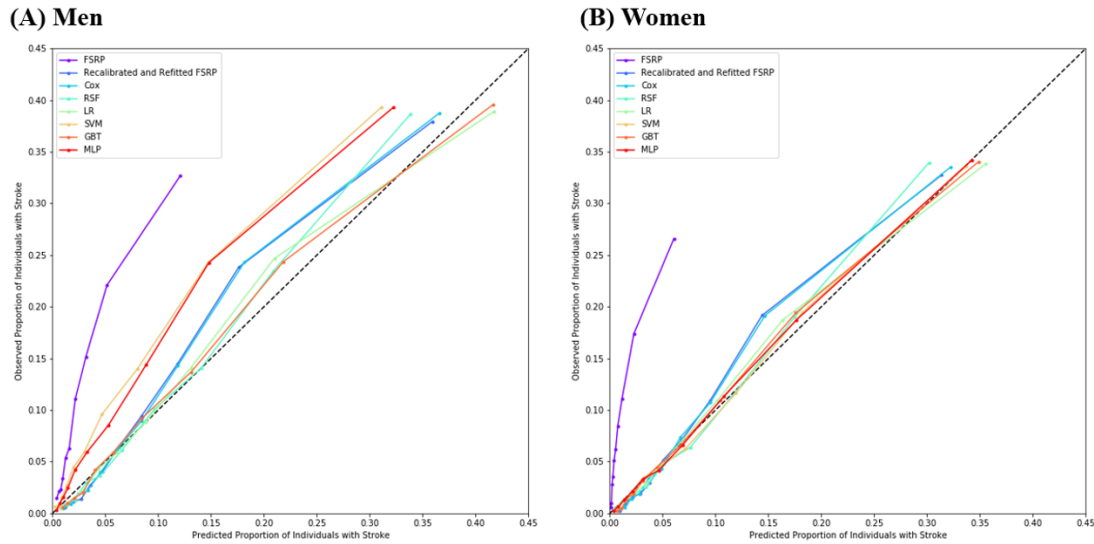


Figure 5.3: Calibration plots for the 2017 Framingham Stroke Risk Profile (FSRP), a recalibrated and refitted FSRP, Cox, random survival forest (RSF), logistic regression (LR), support vector machine (SVM), gradient boosted tree (GBT), and multilayer perceptron (MLP) models in (A) men and (B) women. Each point represents a decile of predicted risk.

Evaluation of risk prediction models at varying intervals of follow-up (0-3 years, 3-6 years, 6-9 years) demonstrated comparable relative performance between models (Figure 5.4, Appendix: Tables C.2-C.4). However, binary classification models (LR, SVM, GBT, and MLP), which were retrained for each prediction task, yielded substantial improvements over survival models (FSRP, Cox, RSF) at all intervals of follow-up and for both sexes (Figures 5.4A and 5.4B). The AUROCs for all models decreased monotonically for later intervals of follow-up. For most models, calibration also declined at later intervals of follow-up, but was less sensitive to

changes in times scales than discrimination performance (Figures 5.4C and 5.4D). At all intervals of follow-up, MLP and SVM (after isotonic regression) had a tendency for poor calibration in men, but were well-calibrated in women. Once again, the FSRP without recalibration and refitting, yielded the worst calibration, substantially underestimating stroke risk at all intervals of follow-up.

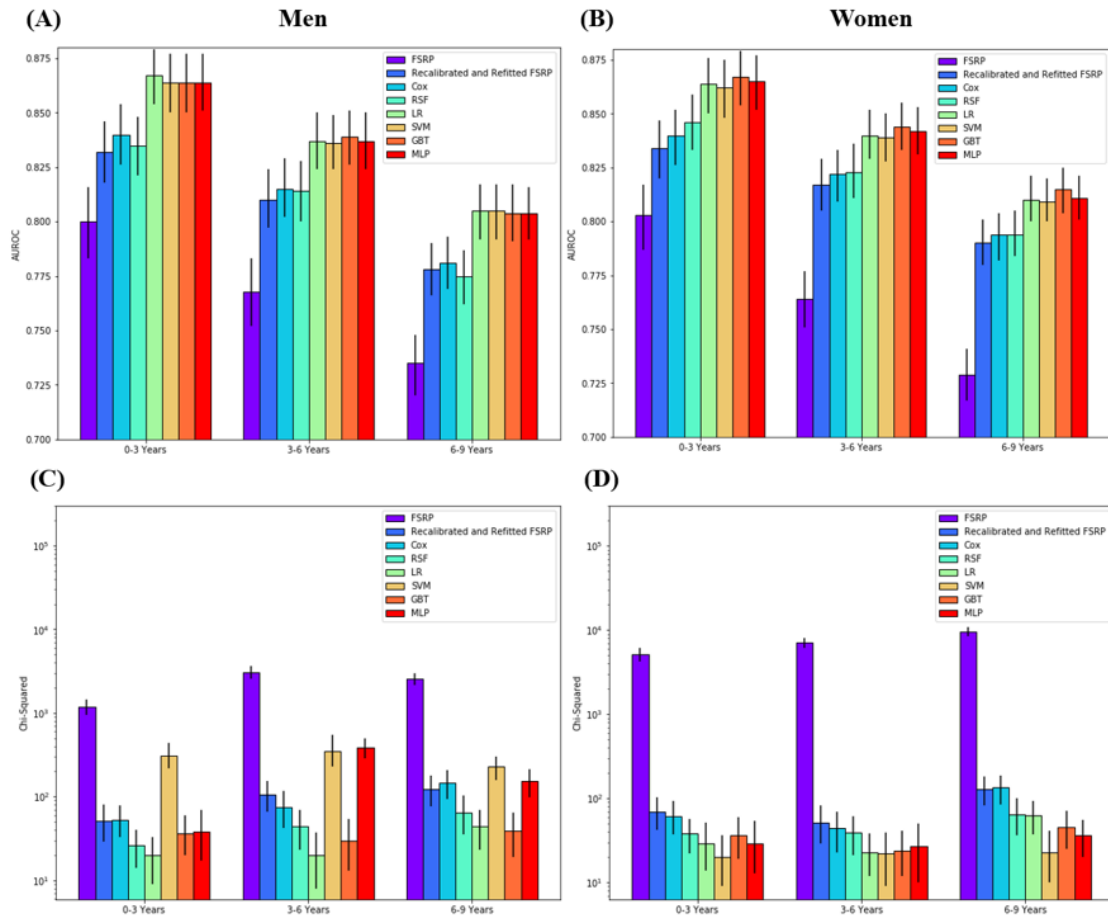


Figure 5.4: Discrimination (subplots A and B) and calibration (subplots C and D) performance in men and women respectively for risk prediction of stroke at various time scales (0-3 years, 3-6 years, 6-9 years after baseline). Comparisons made between the 2017 Framingham Stroke Risk Profile (FSRP), a recalibrated and refitted FSRP, Cox, random survival forest (RSF), logistic regression (LR), support vector machine (SVM), gradient boosted tree (GBT), and multilayer perceptron (MLP) models.

GBT remained among the best-performing of the ML models based on discrimination and calibration metrics. Additional analyses of the most important GBT features for later intervals of follow-up (Table 5.2) indicated comparable performance in men and women with emphasis on risk factor indicators related to

age, blood pressure, physical activity, and geographic area. However, characteristics such as retirement and number of children became relatively more important at later intervals of follow-up (3-6 years and 6-9 years) than at 0-3 years from baseline.

Table 5.2: Top 10 most important GBT features for risk prediction of stroke in men and women from 0-3 years, 3-6 years, and 6-9 years.

Feature Ranking	Men			Women		
	0-3 Years	3-6 Years	6-9 Years	0-3 Years	3-6 Years	6-9 Years
1	Age	Age	Age	Age	Age	Age
2	SBP	SBP	SBP	SBP	Harbin	Harbin
3	Age 65+	Harbin	Harbin	Harbin	SBP	SBP
4	DBP	Age 65+	Age 65+	Age 65+	Age 65+	Age 65+
5	SBP_noHRX	DBP	SBP_HRX	SBP_HRX	SBP_HRX	SBP_HRX
6	Harbin	SBP_HRX	SBP_noHRX	DBP	Children	Children
7	SBP_HRX	SBP_noHRX	Met	SBP_noHRX	Retired	DBP
8	Met	Met	DBP	Met	SBP_noHRX	SBP_noHRX
9	Phys_Activity	HTN_RX	Phys_Activity	Phys_Activity	DBP	Met
10	HTN_diag	Retired	Children	HTN_diag	Haikou	Phys_Activity

SBP: systolic blood pressure, DBP: diastolic blood pressure, Age 65+ is a binary variable (0 if < 65 years old, 1 otherwise), SBP_noHRX: systolic blood pressure if not on blood pressure drugs, SBP_HRX: systolic blood pressure if on blood pressure drugs, Met: daily physical activity (metabolic equivalent), Phys_Activity: daily physical activity (MET-hours), HTN_diag: hypertension diagnosis.

5.3.3 Prediction of Best Model for Risk Screening

After identifying GBT as the best-performing ML model, risk predictions estimated using the Cox and GBT models were used to screen individuals at high-risk of stroke in the validation set in order to train a decision tree to select which model to use for a particular individual, given disagreement about the individual's risk classification. For the purposes of this study, individuals were classified as "high-risk" if they had >10% predicted risk of stroke in 9 years. In cases of disagreement between models, either the Cox or GBT model was identified as the better classifier if it classified an individual with stroke as "high-risk" or classified a stroke-free individual as "not high risk".

The t-SNE visualisations of individuals in the validation set and test set (Figure 5.5) indicated high levels of agreement between both the Cox and GBT models for stroke risk prediction in both men and women. Disagreements occurred in

only 5% of men and women, of whom 10% of men and 12% of women would go on to experience a stroke event.

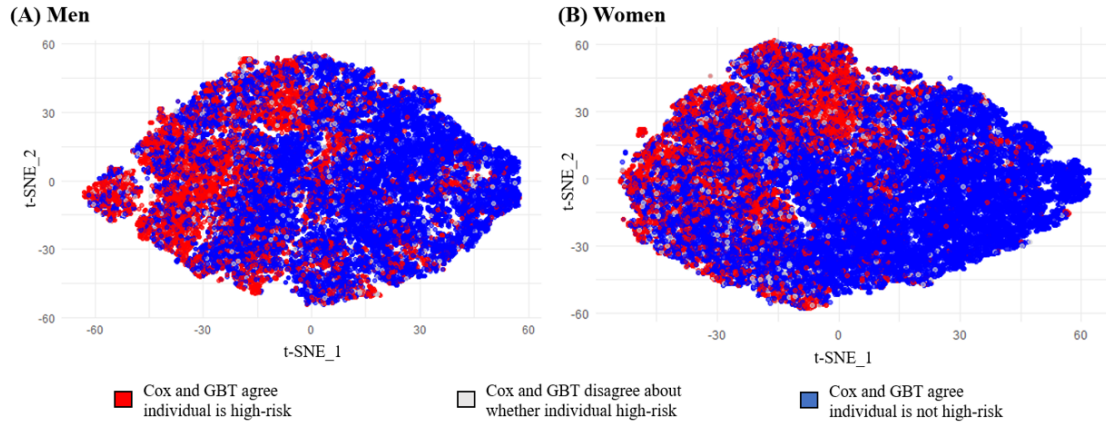


Figure 5.5: t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisations of CKB individuals in validation set and test set. Individuals are coloured by agreement between Cox and GBT risk prediction models for screening of high-risk individuals. High-risk individuals were defined as individuals with $>10\%$ predicted 9-yr risk of stroke. t-SNE plots were created using Rtsne package version 0.15 with perplexity = 50, theta = 0.5, and max iterations = 3000.

Among the 4,281 men and 6,441 women in the test set (after excluding stroke-free, right-censored individuals), sex-specific decision trees were used to resolve 590 disagreements (219 in men, 371 in women) and selected GBT as the better classifier 31% of the time (34% for men, 29% for women). Among individuals in the test set for whom the Cox and GBT models disagreed, the decision trees yielded good discrimination for model selection (AUROC: 0.71 in men, 0.74 in women) and correctly selected the better classifier with an accuracy of 73% in men and 70% in women. The most important features for selecting the better classifier included blood pressure, age, and geographic area-related risk factor indicators for both sexes, in addition to additional features reported in Table 5.3.

5.3.4 Evaluation of Ensemble Approach for Risk Screening

In both men and women, the ensemble approach yielded a higher accuracy (76% in men, 80% in women) and specificity (76% in men, 81% in women) compared with either Cox-only or GBT-only approaches (Table 5.4). The ensemble approach

Table 5.3: Top 10 most important decision tree features for selecting between Cox and GBT for ensemble risk screening approach.

Feature Ranking	Men		Women	
	Feature	Importance Score	Feature	Importance Score
1	SBP	0.3035	Age	0.2053
2	Age 65+	0.1399	SBP_noHRX	0.1164
3	Has health coverage	0.1012	Harbin	0.1045
4	Retired	0.0640	SBP	0.0814
5	Standing height	0.0590	Eats rice daily	0.0802
6	Harbin	0.0454	Hunan	0.0631
7	Waist to hip ratio	0.0355	DBP	0.0510
8	Waist size	0.0339	Father alive	0.0492
9	Age	0.0335	Waist size	0.0359
10	Phys_Activity	0.0315	Never/rarely brushes teeth	0.0315

SBP: systolic blood pressure, DBP: diastolic blood pressure, Age 65+ is a binary variable (0 if <65 years old, 1 otherwise), SBP_noHRX: systolic blood pressure if not on blood pressure drugs, Phys_Activity: daily physical activity (MET-hours). Importance scores calculated as the normalised total reduction of the criterion brought by that feature (Gini importance).

also yielded the highest PPV in men (26%). The GBT-only approach yielded the highest sensitivity (80% in men, 74% in women), and the Cox-only approach did not provide the best performance using any metric. Differences in performance were incremental for all metrics with overlapping 95% confidence intervals. The detailed confusion matrices for each screening approach are reported for both sexes in the Appendix (Table C.5). Good agreement was observed between all models with κ values ranging from 0.85 to 0.96 (Table 5.5).

5.4 Discussion

In this study, involving almost a 100-fold larger population than the Framingham Study [8], we developed novel risk scores for prediction of stroke in a contemporary Chinese cohort. Previous population-based prospective studies [10, 55, 71] highlighted the need for novel risk scores for use in Chinese adults and proposed new Cox-derived models based on these populations. However, the models derived in the present study were based on a substantially larger (5-fold compared to the China-PAR study [10]) and more contemporary population.

Table 5.4: Summary metrics from screening of high-risk individuals in test set using (i) a Cox-only approach, (ii) a GBT-only approach, and (iii) an ensemble approach in which a decision tree selects between Cox and GBT in cases of disagreement over an individual's risk classification.

Metric	Men			Women		
	Cox-Only [95% CI]	GBT-Only [95% CI]	Ensemble [95% CI]	Cox-Only [95% CI]	GBT-Only [95% CI]	Ensemble [95% CI]
Sensitivity	76% [72%-80%]	80% [76%-84%]	76% [72%-80%]	68% [64%-72%]	74% [70%-78%]	67% [64%-71%]
Specificity	75% [74%-77%]	74% [73%-76%]	76% [75%-78%]	80% [79%-81%]	78% [77%-79%]	81% [80%-82%]
PPV	25% [23%-27%]	25% [23%-28%]	26% [23%-28%]	24% [22%-26%]	23% [21%-25%]	24% [22%-26%]
NPV	97% [96%-97%]	97% [96%-98%]	97% [96%-97%]	97% [96%-97%]	97% [97%-98%]	97% [96%-97%]
Accuracy	75% [74%-77%]	75% [74%-76%]	76% [75%-77%]	79% [78%-80%]	77% [76%-78%]	80% [79%-81%]

Table 5.5: Agreement among Cox-only, GBT-only, and ensemble approaches for screening of high-risk individuals in the test set, assessed using Cohen's kappa (κ).

Agreement between:	Men			Women		
	Cox-Only and GBT-Only	Cox-Only and Ensemble	GBT-Only and Ensemble	Cox-Only and GBT-Only	Cox-Only and Ensemble	GBT-Only and Ensemble
Cohen's Kappa κ	0.88	0.96	0.92	0.85	0.95	0.89

In contrast to previous studies of risk prediction of stroke in Chinese adults, we compared both conventional Cox model-based approaches and ML techniques for risk prediction to assess the potential of ML techniques for improved risk prediction. Consistent with findings for cardiovascular risk prediction [72], we demonstrated that ML techniques improved 9-yr risk prediction of stroke over Cox models, with gradient boosted trees (GBT) providing the best discrimination and calibration performance. Improvements over the Cox model were particularly evident for binary classification models that predicted stroke at narrower intervals of follow-up (0-3 years, 3-6 years, and 6-9 years from baseline). This may be due to the fact that survival models such as Cox and random survival forests (RSF) are optimised across the overall 9 year follow-up period, while binary classification models are able to be retrained for optimal performance in each particular time interval of interest.

All models substantially outperformed the 2017 Framingham Stroke Risk Profile, which greatly underestimated stroke risk in CKB.

While the discrimination improvements of ML over Cox models were marginal, such incremental improvements can translate to meaningful population health benefits. For example, a recent analysis of 100,000 UK adults reported that polygenic risk scores for CVD with improvements of just 0.012 in the C-index could help to prevent 7% more CVD events than conventional risk scores alone [117]. Moreover, the substantial calibration improvements of ML approaches such as GBT over Cox models are highly relevant for clinical practice, which relies on absolute risk thresholds for risk screening. Contemporary clinical guidelines recommend using absolute risk predictions from Cox models to screen individuals at high-risk of stroke, who are then prioritised for initiation of drug treatments. For example, the 2013 guidelines of the American College of Cardiology and the American Heart Association [118] recommend initiation of statin therapy for those with a $\geq 7.5\%$ 10-year CVD risk as assessed by the Pooled Cohort Equations, while in the UK, the cutoff is $\geq 10\%$ risk as assessed by QRISK3 [119]. In such settings, underestimation of stroke risk due to poor calibration of models can translate to a failure to identify high-risk individuals who could benefit from statins or other preventative treatments.

Using a threshold of $\geq 10\%$ 9-year stroke risk, we found that an ensemble approach that combined Cox and GBT models had a higher accuracy, specificity, and PPV for stroke prediction than either the Cox-only or GBT-only approaches. However, such improvements were marginal and warrant assessment for reproducibility in external validation studies. We have provided statistical code, available at <https://github.com/ckbiobank/ckb-stroke-risk-models>, to enable others to replicate these findings in other populations.

In contrast with typical ensemble approaches that use voting or averaging of base model outputs, a major strength of our ensemble approach is its default reliance on the Cox model. Cox model-based approaches to risk prediction are widely used in clinical practice, and their relative simplicity and interpretability have been challenges to the adoption of novel ML-based methods [76]. We found that the

component Cox and GBT models failed to agree on risk prediction for stroke in about 5% of individuals, and of these, our proposed ensemble approach selected GBT as the better-performing model about one third of the time. This suggests that, in practice, our ensemble approach would override the risk classification of the Cox model for only a small proportion (1%-2%) of individuals. Meanwhile, clinicians could continue to use the output from Cox model-derived scores, without any loss of predictive performance for the vast majority of individuals at high-risk of stroke. Rather than changing the existing paradigm of stroke risk prediction, our proposed ensemble approach is designed as an incremental change to current practice, which could help to facilitate more widespread use and trust of ML methods for risk prediction in clinical practice [120].

Other barriers to the adoption of complex high-dimensional models for application in clinical practice include the availability of certain risk factor data and the need for regular updating and recalibration of such models. However, as electronic health records (EHRs) become more detailed and widespread, they may mitigate these issues by providing detailed individual-level data and enabling automatic updating and recalibration of complex ML models to local practices [120].

This study had several limitations. First, atrial fibrillation (AF), which is commonly included in risk scores for CVD and stroke [8, 47], was not recorded in the China Kadoorie Biobank (CKB) and could not be included in the models. However, other population-based studies of comparable age groups in China indicated that the prevalence of AF was substantially lower in China than in the Framingham Stroke Risk Profile cohort (1.7% vs 7.1%) [8, 104]. Hence, omission of AF is unlikely to have had a material impact on stroke risk prediction in CKB. Second, the exclusion of right-censored data is an inherent limitation of training the binary classification models presented in this study. Although few participants in CKB were lost to follow up and exclusion of these individuals did not lead to a reduction in model performance, care should be taken when developing similar models in other study populations. Finally, the risk equations outlined in the present report were not designed for immediate implementation in clinical practice. Further

work is needed to validate and refine the proposed risk prediction models and screening approaches from this study in independent populations in China, and potentially other LMICs, since the CKB cohort may not be representative of other target populations. Additional work should also compare the cost and benefits of implementing such approaches over existing care guidelines before implementing them in clinical practice.

5.5 Conclusion

Novel risk scores for stroke have been developed using data from a contemporary cohort of 0.5 million Chinese adults. Use of machine learning (ML) techniques improved risk prediction over traditional Cox model approaches, with gradient boosted trees (GBT) providing the best discrimination and calibration performance. An ensemble approach was also proposed to screen for individuals at high-risk of stroke who may benefit from more intensive treatment. The ensemble approach identified high-risk individuals with marginal improvements to accuracy, specificity, and positive predictive value (PPV) over either Cox or GBT models alone. By identifying a small portion of individuals who would benefit from ML predictions, our ensemble approach provides an incremental benefit beyond current clinical practice that has potential to translate to important population health benefits and facilitate the adoption of ML-based risk calculators in clinical practice.

6

Single v. Sequential Risk Factor Measurements for Stroke Risk Prediction

6.1 Introduction

As previously discussed, current guidelines for primary prevention of stroke advocate use of risk prediction equations, including the Pooled Cohort Equations [46] for cardiovascular disease (CVD) or the Framingham Stroke Risk Profile [8] for stroke [30, 31]. Such risk scores are typically derived using Cox proportional hazards models that generate risk predictions based on measurements of risk factors recorded at a single baseline visit. In Chapter 4, we evaluated the utility of Cox models for stroke risk prediction in the China Kadoorie Biobank (CKB), and in chapter 5, we compared Cox models to a range of machine learning (ML) approaches for risk prediction based on a single measurement of risk factors recorded at a baseline visit.

However, adults without established CVD are recommended to have their risk reassessed every 3 to 5 years [46, 121, 122], and it remains unclear if CVD prediction can be improved using repeated measurements of time-dependent risk factors including blood pressure (BP), cholesterol levels, body weight, and lifestyle factors. BP has been extensively studied, and while longitudinal indices of BP appear to have incrementally stronger associations with CVD than “present” BP, head-to-head comparisons of their predictive utility for incident disease are limited [123].

Previous studies comparing the predictive utility of sequential versus single risk factor measurements have reported conflicting results, ranging from no improvement [124, 125] to statistically significant incremental improvements (Δ C-index: 0.0023-0.0040) [126] or substantial improvements (Δ C-index: 0.05-0.06) [127, 128], despite using similar modelling approaches. Moreover, none of these previous studies examined major risk factors for incident stroke or involved the Chinese population, in which incidence of stroke is particularly high.

Consequently, substantial uncertainty persists about the relevance of sequential measurements of established CVD risk factors recorded on two or three separate visits versus those recorded on a single visit for prediction of incident stroke in Chinese adults. The aims of the study presented in this chapter were: (i) to examine changes in CVD risk factors and 3-year absolute risks of stroke in Chinese adults from 2004-2014, and (ii) to compare the utility of sequential versus single measurements of CVD risk factors for prediction of stroke in a contemporary study of Chinese adults.

6.2 Methods

6.2.1 Inclusion Criteria and Data Preprocessing

Details of the design and methods used in the China Kadoorie Biobank (CKB) are reported in detail in Chapter 3. In contrast to Chapters 4 and 5, the analyses presented in this chapter utilised data from the two CKB resurveys (conducted in 2008 and 2013-2014, respectively) in addition to the baseline survey conducted from 2004-2008.

In CKB, 14,999 individuals had data available from three separate visits at baseline, first resurvey, and second resurvey (Figure 6.1). Among these individuals, there were 1,760 incident strokes recorded before the end of follow-up.

For some individuals included at second resurvey ($n=97$), only physical measurements were recorded without a full questionnaire administered. After removing these individuals, 14,902 individuals remained. Among these individuals, there were 1,733 incident strokes recorded before the end of follow up.

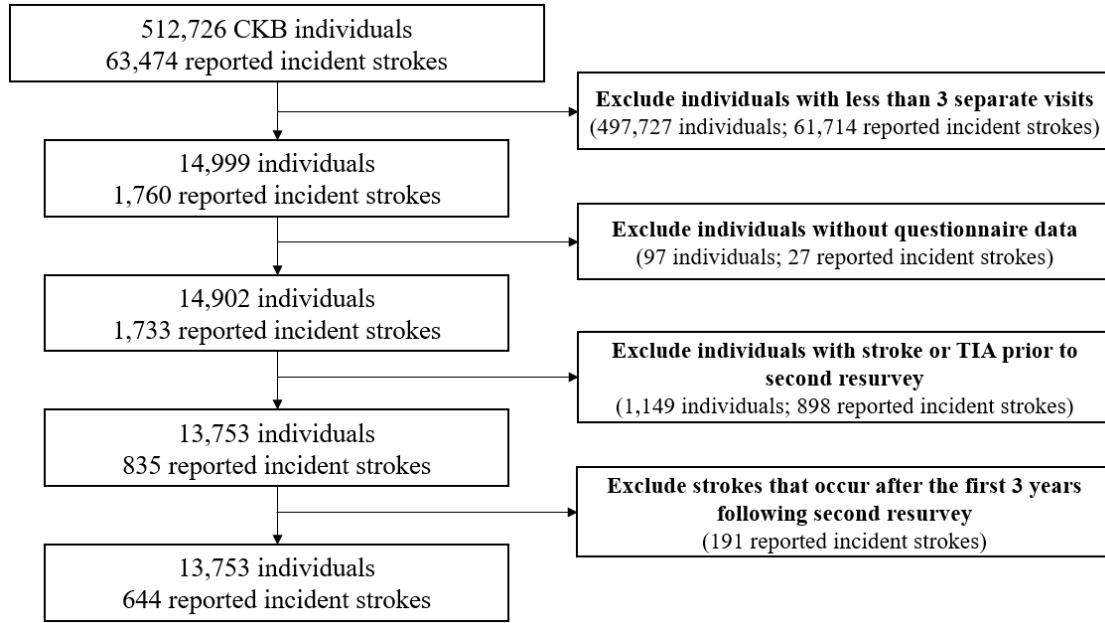


Figure 6.1: Inclusion criteria for analyses presented in Chapter 6.

We excluded all individuals with stroke prior to second resurvey (either self-reported prior stroke or TIA at any visit, or follow-up data indicating a stroke event prior to the second resurvey). After removing these individuals ($n=1,149$), 13,753 individuals remained. Among these individuals, there were 835 incident strokes recorded before the end of follow up.

Finally, after considering only incident first stroke cases that occurred within 3 years from the second resurvey, we ended up with 13,753 included individuals (5,152 men; 8,601 women) with 644 first incident stroke cases (267 in men; 377 in women) recorded within three years following the second resurvey. The 3-year follow-up period was selected since some participants completed the second resurvey as late as 2014, with global end of follow-up being January 1, 2018.

6.2.1.1 Maintaining Consistency Between Visits

In order to maintain consistency between risk factor inputs recorded at each visit, we considered only those that were recorded at all three surveys (i.e., baseline, first resurvey, and second resurvey). Region (including urban/rural classification) and sex were the only exceptions. Region and sex were recorded at the baseline survey and assumed to remain unchanged throughout the study.

Household income was recorded with additional high-income bands in the second resurvey, but for the purposes of the present study, were standardised to match the wider income bands used in the baseline questionnaire and first resurvey.

Self-reported diagnoses of acute myocardial infarction, angina, and other ischaemic heart disease were recorded separately at second resurvey, but were grouped together to match the variable for self-reported diagnoses of coronary heart disease at baseline and first resurvey.

Self-reported diagnoses of emphysema and bronchitis were recorded separately at second resurvey, but were grouped together to match the single variable for self-reported diagnoses of emphysema OR bronchitis at baseline and first resurvey.

Self-reported diagnoses of depression, anxiety, and other psychological disorders were recorded separately at second resurvey, but were grouped together to match the single variable for psychological disorders at baseline and first resurvey.

6.2.1.2 Training and Test Sets

Included CKB individuals were randomly divided into a training set (85%) and test set (15%) for model development and internal validation, respectively.

For survival models such as the Cox proportional hazards models developed in this study, right-censored individuals were included in the training set. The survival model training set included 11,689 individuals (4,379 men; 7,310 women) with 547 incident first stroke cases during the 3-year follow-up period (227 in men; 320 in women). For binary prediction models such as the gradient-boosted tree models (GBTs) developed in this study, right-censored individuals who were lost to follow-up were excluded in the training set (since it was unknown whether or not they had a stroke during the full follow-up period). The training set for these models included 11,464 individuals (4,259 men; 7,205 women) with 547 incident first stroke cases during the 3-year follow-up period (227 in men; 320 in women). Right-censored individuals were further excluded from the test set for all models since it was unknown if they could have suffered a stroke within the time interval of interest.

The test set included 2,017 individuals (751 men; 1,266 women) with 97 incident first stroke cases during the 3-year follow-up period (40 in men; 57 in women).

6.2.1.3 Missing Values

Missing values in the CKB baseline questionnaire were uncommon and have been previously described in Section 3.4.2. Similarly, missing values were uncommon in the first resurvey and second resurvey, consisting of less than 0.03% of all risk factor input data.

Considering measurements of the same risk factor input at separate visits as distinct risk factor input variables, only 80 of the approximately 400 available risk factor input variables had any missing values. Of the 80 risk factor input variables with missing values, there was a maximum of only 60 missing values for 79 of these features. Most of these variables corresponded to family medical history information. 10 individuals were also missing self-reported personal medical history information in the second resurvey. Fat percentage at second resurvey was the only outlier, having 166 missing values across the 13,753 included participants.

To address missing values for family medical history, 4 binary risk factor input variables were included to indicate missing medical history for the individual's mother, father, siblings, or children, respectively, at each of the three visits. All missing values were imputed using sex-specific mean values from the training set.

6.2.2 Model Development

6.2.2.1 Cox Models

To reflect the Cox models currently used in clinical practice, such as the Pooled Cohort Equations [46] and the Framingham Stroke Risk Profile [8], we first developed regularised and region-stratified, sex-specific Cox models for 3-year risk prediction of stroke using data on risk factors recorded at the second resurvey. These Cox models utilise risk factors measurements recorded at a single visit for prediction of future stroke events and serve as an established gold standard against which other modelling approaches can be compared.

Cox models were derived separately for men and women with stratification by CKB area (5 urban and 5 rural regions). For each model, separate baseline survival functions were developed for each CKB area. The baseline hazard function (and corresponding baseline survival function) were derived from the non-parametric Breslow estimator [84, 85]. 10-fold cross-validated least absolute shrinkage and selection operator (LASSO) regularisation was used within the training set for selecting a subset of risk factor inputs from all CKB variables and yielded slightly different numbers of selected risk factor inputs (66 risk factor inputs for men, 70 risk factor inputs for women). The specifics of the variable selection process we implemented have been previously described [86, 87].

6.2.2.2 Linear Mixed Effects Models

In order to analyse temporal trends in absolute risks of stroke among Chinese individuals from 2004-2014, the Cox models (developed based on second resurvey data only) were retrospectively applied, for each individual, to risk factor measurements also recorded at the baseline and first resurvey. This yielded a total of three separate risk estimates for each individual, with each estimate based on risk factor measurements recorded at a single visit. Population-level and individual-level trends in the risk estimates were then analysed using linear mixed effects models (LMEs).

LMEs with individual-level random slopes and random intercepts were developed separately for each sex to analyse temporal trends in the Cox-generated 3-year stroke risk estimates for each individual. Intercepts were representative of predicted 3-year stroke risk at second resurvey while slopes were representative of annual changes in the 3-year predicted stroke risk in the decade preceding the second resurvey.

The form of the fitted LMEs was specified as:

$$StrokeRisk_{ij} = (b_{0j} + \beta_0) + (b_{1j} + \beta_1)t_{ij} + \epsilon_{ij} \quad (6.1)$$

$$b_{0j} \sim \mathcal{N}(0, \tau_1^2) \quad (6.2)$$

$$b_{1j} \sim \mathcal{N}(0, \tau_2^2) \quad (6.3)$$

$$\text{Cov}(b_{0j} b_{1j}) = \tau_{12} \quad (6.4)$$

where $StrokeRisk_{ij}$ denotes the 3-year stroke risk for visit i and individual j , β_0 represents the population-level fixed intercept, β_1 represents the population-level fixed slope, b_{0j} represents the random intercept for individual j , b_{1j} represents the random slope for individual j , t_{ij} represents the time before second resurvey (in years), and ϵ_{ij} represents the random error.

In order to make predictions for a previously unseen individual, a maximum likelihood estimate of the individual's random slopes and random intercepts were determined based on the posterior distribution conditioned on the individual's predicted risks and the trained LME parameters. These features, along with other longitudinal features and risk factor input data, were provided to GBTs for risk prediction, as described below.

LMEs were derived using the `lmer()` function of the `lme4` package [129] in R version 3.6.1.

6.2.2.3 Gradient Boosted Tree Models

Finally, to assess the value of machine learning (ML) approaches and sequential measurements of risk factor inputs for 3-year prediction of stroke, we developed sex-specific gradient boosted tree (GBT) models using four modelling approaches and compared their performance against the Cox models. GBT was chosen because of its ability to model non-linear interactions between risk factor inputs, incorporate longitudinal features, and achieve best-in-class performance for stroke risk prediction in other recent studies [128, 130].

In the first GBT modelling approach, we used only single measurements of risk factor inputs recorded at the second resurvey in order to provide a direct comparison against the Cox models, which were developed using the same input data.

In the second GBT modelling approach, we used the raw sequential measurements of risk factor inputs recorded at all three visits (i.e., baseline, first resurvey, second resurvey) with no further feature engineering.

In the third GBT modelling approach, we engineered features to provide a longitudinal summary of the sequential measurements recorded at all three visits,

in a similar manner reported by Cho *et al.* [127] and Zhao *et al.* [128]. Namely, we provided the models with (i) mean, standard deviation, minimum, and maximum values recorded for continuous risk factor inputs, and (ii) mean and standard deviation of recorded values for categorical risk factor inputs.

In the fourth GBT modelling approach, we engineered features to provide a longitudinal summary of the Cox-generated risk estimates at all three visits and combined these with single measurements of risk factor inputs recorded at the second resurvey. This fourth modelling approach was adapted from two-stage modelling approach described by Sweeting *et al.* [124], and the engineered features included (i) individual-level random slopes and random intercepts from the linear mixed effects models, (ii) variance of Cox-generated risk estimates across the three visits, and (iii) absolute changes in Cox-generated risk estimates between consecutive visits. While the third modelling approach incorporated longitudinal summaries of the individual risk factor inputs, the fourth modelling approach provided a longitudinal summary of the overall stroke risk estimates for each individual.

All GBT hyperparameters were tuned to maximise AUC using repeated 3-fold cross-validated grid searches within the training set, and all GBT models were trained and calibrated with isotonic regression using 3-fold cross-validation. Additional details of the hyperparameter tuning process for the GBT models are provided in Appendix D.1.

6.2.3 Statistical Analysis

All models were evaluated for predictive performance using the test set, in which we compared the performance of the Cox and GBT models for predicting stroke during the 3 years following the second resurvey. Model discrimination and calibration were assessed using area under the receiver operating characteristic curve (AUC) and χ^2 statistics from the Hosmer-Lemeshow (for binary classification models) and Nam-D’Agostino (for survival models) tests. Confidence intervals for AUCs were constructed using 1,000 bootstrapped samples from the test set. Better risk discrimination was indicated by higher AUCs, and better calibration was

indicated by lower χ^2 values. To assess the changes in predictive performance of each modelling approach against the Cox models, we performed Delong’s test [131] for Δ AUCs and calculated category-free net reclassification index (NRI) and integrated discrimination improvement (IDI). Statistical analyses were performed using Python version 3.7.0 and R version 3.6.1.

6.3 Results

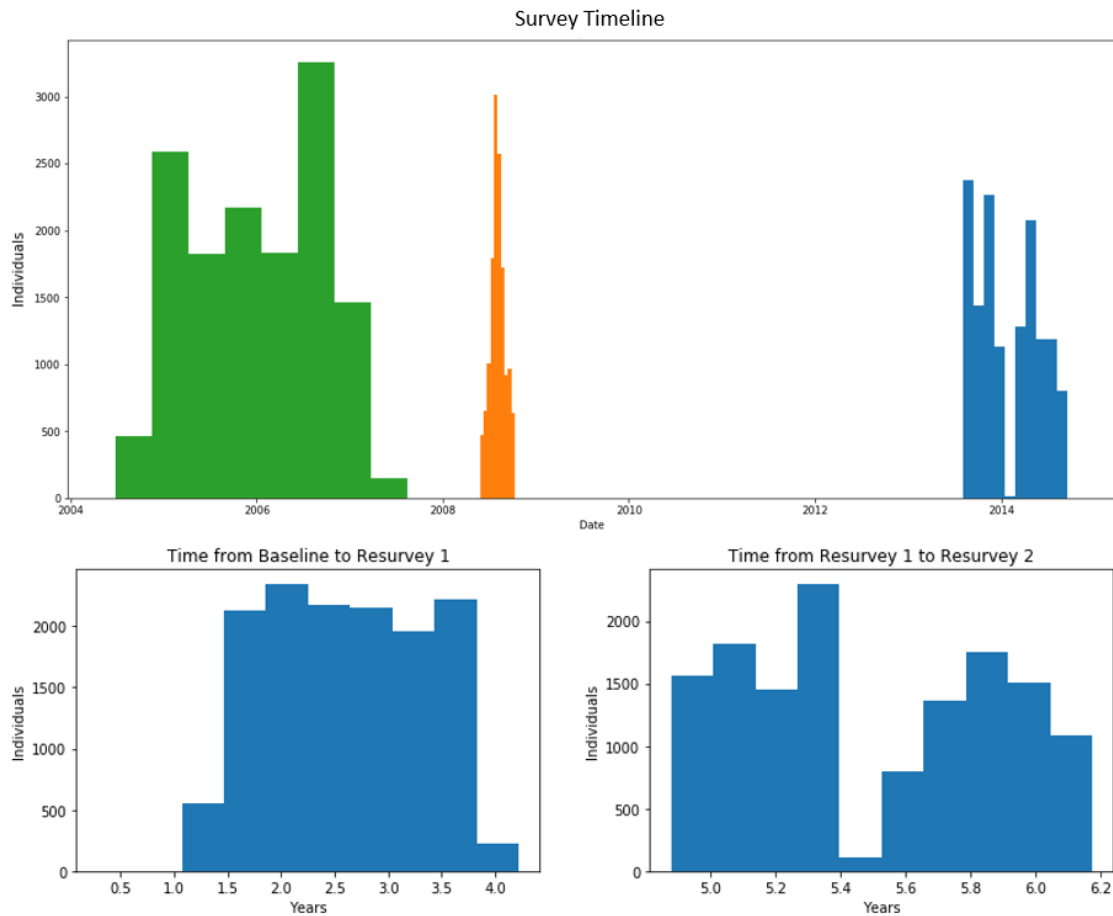


Figure 6.2: Timing of baseline questionnaire, resurvey 1, and resurvey 2 for included CKB individuals.

Details about the timing of risk factor measurements are provided in Figure 6.2. Among the 13,753 CKB participants (5,152 men; 8,601 women), 644 incident cases of first-ever stroke (267 in men; 377 in women) were recorded within the first three years following the second resurvey. Compared with those who had no

stroke, individuals who had a first stroke were older and more likely to have prior history of coronary heart disease (CHD), diabetes, or hypertension (Table 6.1). In both men and women, age, prior history of coronary heart disease, diabetes, use of antihypertensive medication, and mean levels of systolic blood pressure increased from baseline to second resurvey, while the prevalence of current smoking decreased. Further details about the time-evolution in distributions of well-established stroke risk factors are provided in the Appendix (Figures D.1-D.6).

Table 6.1: Distribution of major risk factors for stroke in individuals at baseline (2004-2008), 1st resurvey (2008), and 2nd resurvey (2013-2014) in the China Kadoorie Biobank.

Men						
Risk factor	Baseline (2004-2008)		Resurvey 1 (2008)		Resurvey 2 (2013-2014)	
	No	Total	No	Total	No	Total
	Stroke	Stroke	Stroke	Stroke	Stroke	Stroke
	(n=4,885)	(n=267)	(n=4,885)	(n=267)	(n=4,885)	(n=267)
Age, mean, yr	51.3	60.3	53.9	62.9	59.3	68.4
Current smoking, %	68.7	61.4	65.8	58.1	55.9	50.4
CHD, %	1.8	7.9	2.8	9.7	5.9	13.9
Age ≥ 65 yrs, %	11.5	33.7	16.2	46.4	29.3	64.4
Diabetes, %	3.5	7.1	5.6	13.1	9.3	16.9
BP-lowering treatment, %	6.6	12.7	9.7	21.3	18.9	27.3
SBP, mean, mmHg	132	140	126	134	135	147
Women						
Risk factor	Baseline (2004-2008)		Resurvey 1 (2008)		Resurvey 2 (2013-2014)	
	No	Total	No	Total	No	Total
	Stroke	Stroke	Stroke	Stroke	Stroke	Stroke
	(n=8,224)	(n=377)	(n=8,224)	(n=377)	(n=8,224)	(n=377)
Age, mean, yr	50.1	57.1	52.7	59.8	58.2	65.3
Current smoking, %	2.9	4.2	2.5	3.2	2.2	2.4
CHD, %	2.3	6.1	3.4	9.5	6.7	17.2
Age ≥ 65 yrs, %	7.9	23.6	12.0	30.2	24.1	50.9
Diabetes, %	4.1	10.3	7.0	16.2	10.8	20.7
BP-lowering treatment, %	8.8	15.4	11.6	20.4	20.7	32.6
SBP, mean, mmHg	130	140	124	133	135	147

Note: “No Stroke” columns include individuals who remained stroke-free until being censored, even if lost to follow-up before 3 years.

According to the Cox models, these temporal risk factor trends yielded increasing 3-year predicted stroke risks separately in individuals with and without stroke (Figure 6.3). On average, 3-year predicted stroke risks increased by 0.3% per annum in men and 0.2% per annum in women, but these risks differed substantially between

individuals (Figure 6.3). Among those with stroke, predicted risk increased 1.0% per annum in men and 0.8% per annum in women, with average predicted risk at second resurvey being 13.9% in men and 12.0% in women. In contrast, among those without stroke, predicted risk increased only 0.3% per annum in men and 0.2% per annum in women, with average predicted risk at second resurvey being 4.3% in men and 3.7% in women. In both sexes, the predicted stroke risk at second resurvey was highly correlated with the annual increase of risk for each individual over the previous decade (men: $r = 0.91$, women: $r = 0.89$; Figure 6.4).

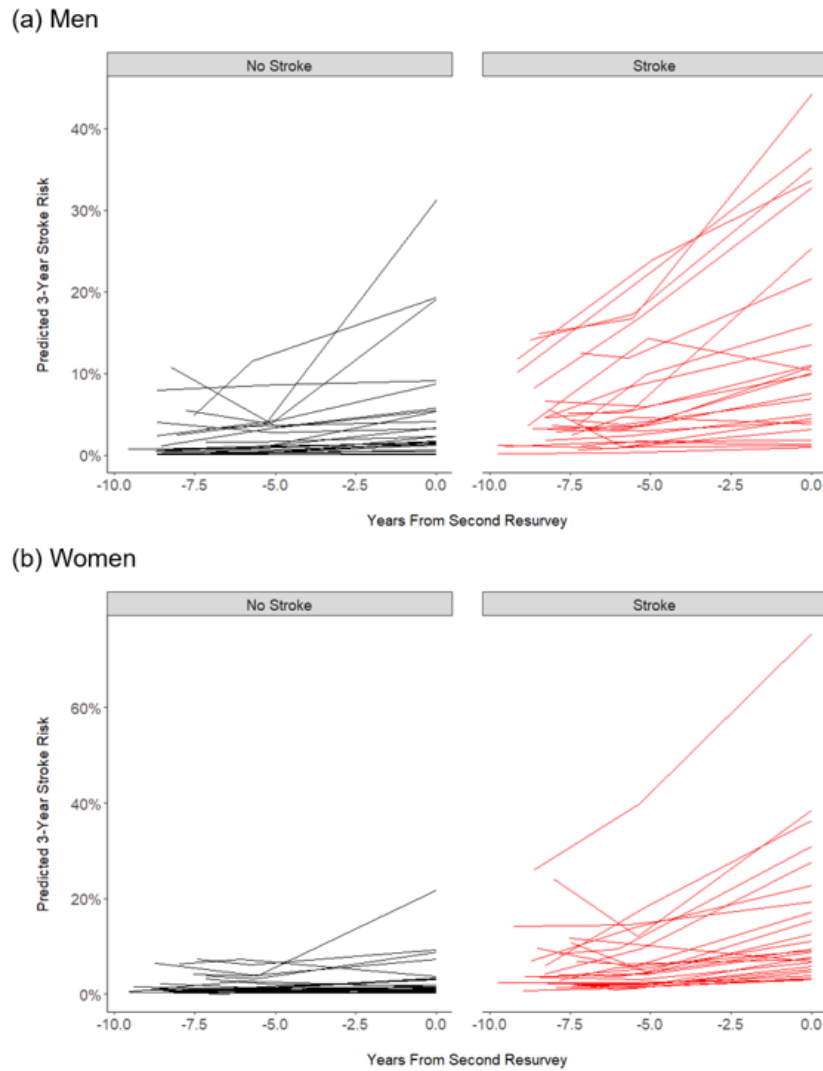


Figure 6.3: Temporal trends in predicted 3-year stroke risk for a random sample of (a) 50 men and (b) 50 women from the China Kadoorie Biobank, stratified by incident stroke outcomes.

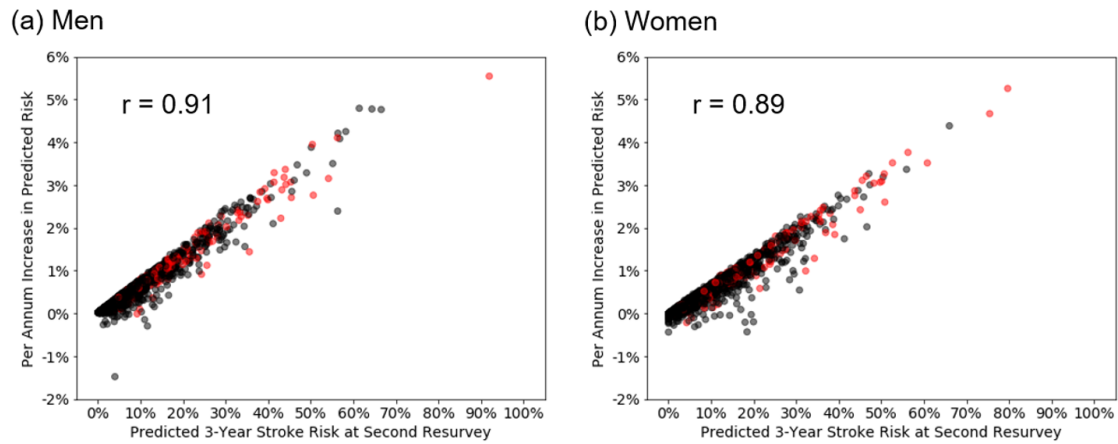


Figure 6.4: Relationship between predicted 3-year stroke risk at second resurvey and per annum increase in predicted risk (fixed effect slope + random effect slope from linear mixed effects model) for (a) men and (b) women in the China Kadoorie Biobank. Black points: individuals with no stroke; Red points: individuals with stroke.

The Cox models, which estimated risk of stroke based on a single measurement of risk factors at second resurvey, predicted incident stroke in the subsequent three years with good discrimination (AUC: 0.779 in men, 0.756 in women) and calibration (χ^2 : 16.8 in men, 17.3 in women) (Table 6.2). Overall, the GBT models yielded slightly higher AUCs in men (0.795-0.811) and lower AUCs in women (0.724-0.750) than the Cox models (Table 6.2). However, none of these differences were statistically significant (Table 6.3) and there was substantial variability across bootstrapped samples. Except for the fourth GBT modelling approach (longitudinal summary of stroke risk estimates at three visits + single measurement of risk factor inputs at most recent visit), the GBT models generally appeared to be better calibrated than the Cox models in both men and women (Table 6.2; Appendix: Figures D.7-D.11). The NRI and IDI values were negative for almost all GBT models. However, once again, the confidence intervals of these values typically included zero (no difference), indicating no significant change in their predictive utility.

Rankings of the top 10 most important features with corresponding Gini importance scores for all GBT models are provided in Tables 6.4-6.5. Across all models, age, blood pressure variables, measurements of size and weight, exercise, and region were identified as important features. However, in the fourth GBT modelling approach, in which a longitudinal summary of the Cox-risk estimates

Table 6.2: Comparison of discrimination and calibration performance for prediction of 3-year risk of stroke using Cox and Gradient Boosted Tree (GBT) modelling approaches.

Modelling Approach and Included Data	Men		Women	
	Discrimination	Calibration	Discrimination	Calibration
	AUC [95% CI]	χ^2 (p-value)	AUC [95% CI]	χ^2 (p-value)
<i>Single measurement of risk factor inputs</i>				
Cox: Single measurement at most recent visit	0.779 [0.709-0.845]	16.8 (p=0.05)	0.756 [0.692-0.814]	17.3 (p=0.04)
GBT: Single measurement at most recent visit	0.811 [0.753-0.867]	5.6 (p=0.78)	0.743 [0.681-0.798]	7.3 (p=0.61)
<i>Sequential measurements of risk factor inputs</i>				
GBT: Sequential measurements at three visits	0.795 [0.721-0.858]	1.9 (p=0.99)	0.741 [0.677-0.796]	14.3 (p=0.11)
GBT: Longitudinal summary ^a of sequential measurements at three visits	0.789 [0.721-0.851]	5.0 (p=0.83)	0.724 [0.660-0.782]	9.2 (p=0.42)
GBT: Longitudinal summary of stroke risk estimates ^b at three visits + Single measurement at most recent visit	0.786 [0.719-0.851]	29.5 (p<0.01)	0.750 [0.683-0.811]	20.2 (p=0.02)

^a Longitudinal summary of risk factor inputs included (i) mean, standard deviation, minimum, and maximum values recorded for continuous risk factor inputs, and (ii) mean and standard deviation for categorical risk factor inputs.

^b Longitudinal summary of risk estimates included individual-level random slopes and random intercepts from linear mixed effects models; variance of Cox-generated risk estimates across three visits; and absolute changes in Cox-generated risk estimates between consecutive visits.

were provided as inputs, these longitudinal features had higher Gini importance scores than any individual risk factor inputs across all of the GBT models.

6.4 Discussion

The present report highlighted temporal changes in mean levels of major stroke risk factors and 3-year absolute risks of stroke in 13,753 geographically-diverse Chinese adults, including 644 incident cases of first stroke occurring between 2004 and 2014. Among men and women who were stroke-free until the second resurvey (2013-2014), the predicted absolute risks of stroke throughout the previous decade increased with each consecutive year of follow-up, together with increasing numbers of individuals with coronary heart disease and diabetes, and increased use of blood pressure-lowering medication. Despite wide variations between individuals, the

Table 6.3: Δ AUC, net reclassification index (NRI) and integrated discrimination index (IDI) of Cox and Gradient Boosted Tree (GBT) modelling approaches for predicting 3-year risk of stroke.

Modelling Approach and Included Data	Men			Women		
	Δ AUC (p-value)	NRI [95%CI]	IDI [95%CI]	Δ AUC (p-value)	NRI [95%CI]	IDI [95%CI]
<i>Single measurement of risk factor inputs</i>						
Cox: Single measurement at most recent visit	Referent	Referent	Referent	Referent	Referent	Referent
GBT: Single measurement at most recent visit	+0.032 (p=0.08)	+0.14 [-0.17,0.44]	-0.03 [-0.08,0.01]	-0.013 (p=0.59)	-0.25 [-0.51,0.01]	-0.03 [-0.06,-0.01]
<i>Sequential measurements of risk factor inputs</i>						
GBT: Sequential measurements at three visits	+0.016 (p=0.57)	-0.20 [-0.52,0.12]	-0.03 [-0.09,0.02]	-0.015 (p=0.42)	-0.13 [-0.39,0.14]	-0.03 [-0.06,-0.01]
GBT: Longitudinal summary ^a of sequential measurements at three visits	+0.010 (p=0.77)	-0.13 [-0.45,0.18]	-0.05 [-0.11,0.01]	-0.032 (p=0.14)	-0.29 [-0.55,-0.03]	-0.04 [-0.06,-0.01]
GBT: Longitudinal summary of stroke risk estimates ^b at three visits + Single measurement at most recent visit	+0.007 (p=0.38)	-0.02 [-0.34,0.30]	0.00 [-0.02,0.02]	-0.006 (p=0.32)	+0.15 [-0.11,0.42]	0.00 [-0.01,0.02]

^a Longitudinal summary of risk factor inputs included (i) mean, standard deviation, minimum, and maximum values recorded for continuous risk factor inputs, and (ii) mean and standard deviation for categorical risk factor inputs.

^b Longitudinal summary of risk estimates included individual-level random slopes and random intercepts from linear mixed effects models; variance of Cox-generated risk estimates across three visits; and absolute changes in Cox-generated risk estimates between consecutive visits.

annual rate of increase of predicted stroke risk was about 3 to 4-fold greater for individuals with stroke compared with those without stroke in the 3 years following the second resurvey. These results suggest that temporal trends in stroke risk can be informative for prediction of incident stroke events.

Nevertheless, while temporal changes in risk are informative for prediction of stroke, the sequential measurements provided limited additional value for prediction compared with risk factors recorded at a single visit. Because of the high correlation

Table 6.4: Top 10 most important features for GBT models in men.

Importance Ranking	GBT: Single measurement at most recent visit		GBT: Sequential measurements at three visits		GBT: Longitudinal summary of sequential measurements at three visits		GBT: Longitudinal summary of stroke risk estimates at three visits + Single measurement at most recent visit	
	Feature	Gini Importance	Feature	Gini Importance	Feature	Gini Importance	Feature	Gini Importance
1	Age @ resurvey 2	0.093	Age @ resurvey 1	0.079	Age @ resurvey 2	0.122	Random intercept from LME	0.213
2	SBP per 10 mmHg, if no HRx @ resurvey 2	0.051	Age @ baseline	0.075	max (SBP mean)	0.036	Random slope from LME	0.163
3	SBP mean @ resurvey 2	0.047	Age @ resurvey 2	0.075	min (sitting height)	0.030	Variance of risk estimates	0.145
4	met @ resurvey 2	0.043	SBP per 10 mmHg, if no HRx @ resurvey 2	0.057	SD (DBP mean)	0.028	Cox risk estimate @ resurvey 2	0.142
5	weight kg @ resurvey 2	0.042	SBP mean @ resurvey 2	0.053	max (DBP mean)	0.027	Change in risk estimate from resurvey 1 to 2	0.087
6	standing height @ resurvey 2	0.039	DBP mean @ resurvey 2	0.028	mean (SBP mean)	0.021	Change in risk estimate from baseline to resurvey 1	0.048
7	DBP mean @ resurvey 2	0.038	waist_cm @ baseline	0.025	mean (children)	0.021	Age @ resurvey 2	0.028
8	sitting height @ resurvey 2	0.038	standing height @ resurvey 2	0.025	Haikou	0.020	standing height @ resurvey 2	0.018
9	BMI @ resurvey 2	0.034	SBP mean @ baseline	0.024	mean (met_hours)	0.020	SBP mean @ resurvey 2	0.016
10	Heart rate @ resurvey 2	0.033	DBP mean @ baseline	0.023	SD (fat percent)	0.018	DBP mean @ resurvey 2	0.015

Table 6.5: Top 10 most important features for GBT models in women.

Importance Ranking	GBT: Single measurement at most recent visit		GBT: Sequential measurements at three visits		GBT: Longitudinal summary of sequential measurements at three visits		GBT: Longitudinal summary of stroke risk estimates at three visits + Single measurement at most recent visit	
	Feature	Gini Importance	Feature	Gini Importance	Feature	Gini Importance	Feature	Gini Importance
1	Age @ resurvey 2	0.165	Age @ baseline	0.046	Age @ resurvey 2	0.068	Cox risk estimate @ resurvey 2	0.217
2	waist_cm @ resurvey 2	0.078	Age @ resurvey 1	0.044	Harbin	0.043	Random intercept from LME	0.216
3	DBP mean @ resurvey 2	0.069	Age @ resurvey 2	0.042	mean (children)	0.028	Random slope from LME	0.159
4	children @ resurvey 2	0.061	met @ resurvey 1	0.037	mean (met_hours)	0.027	Variance of risk estimates	0.135
5	fat_percent @ resurvey 2	0.059	Harbin	0.030	min(children)	0.025	Change in risk estimate from resurvey 1 to 2	0.109
6	SBP per 10 mmHg, if no HRx @ resurvey 2	0.057	SBP per 10 mmHg, if no HRx @ resurvey 2	0.029	mean (DBP mean)	0.020	Change in risk estimate from baseline to resurvey 1	0.019
7	SBP mean @ resurvey 2	0.055	waist_cm @ resurvey 2	0.028	mean (SBP mean)	0.020	fat_percent @ resurvey 2	0.017
8	standing height @ resurvey 2	0.050	waist_cm @ resurvey 1	0.026	mean (SBP per 10 mmHg, if no HRx) max	0.020	Age @ resurvey 2	0.014
9	met_hours @ resurvey 2	0.042	DBP mean @ resurvey 2	0.025	(waist_cm)	0.019	waist_cm @ resurvey 2	0.012
10	Harbin	0.042	DBP mean @ baseline	0.023	mean (met)	0.019	sitting height @ resurvey 2	0.009

between an individual's predicted risk at second resurvey and their per annum increase in risk, the predicted risk at the most recent visit was informative of an individual's temporal risk profile prior to that visit. The individuals at highest risk for stroke at the second resurvey (2013-2014) were largely the same individuals with the fastest growing risk throughout the previous decade. This correlation is an established phenomenon in epidemiology referred to as the "horse-racing effect" [132], by analogy to the fact that horses leading a race at any given moment are those which have been running the fastest up to that point. Thus, the incorporation of temporal information from sequential risk factor measurements did not yield any improved risk prediction compared to models developed using single measurements of risk factor inputs. The results of the present study confirm the utility of single rather than sequential measurements for risk prediction of stroke, as currently recommended in clinical practice [31, 121].

Other studies [124–128] have investigated the utility of sequential risk factor measurements for improving prediction of CVD, but they have not focused specifically on prediction of stroke or risk assessment in a Chinese population. Consistent with the findings of the present study, comparable analyses of approximately 13,000 individuals from the Atherosclerosis Risk in Communities (ARIC) study reported no improvements in AUC/C-index for 3-year risk prediction of CVD [124] (albeit with large standard errors due to a small number of events), or for prediction of coronary heart disease [125], despite exploring a variety of modelling approaches. A larger study involving almost 200,000 participants from the Emerging Risk Factors Collaboration explored the inclusion of repeated risk factor measurements for 5-year CVD prediction, and reported statistically significant but marginal improvements of C-index ranging from 0.0023 to 0.0040 [126]. Meanwhile, a study of 0.5M Korean adults found more substantial improvements in CVD prediction (Δ C-index: 0.05–0.06) by including a longitudinal summary of sequential risk factor measurements recorded at periodic health screenings [127], and a study using electronic health records data of 109,490 individuals from the Vanderbilt University Medical Center

reported improvements in AUC of 0.05-0.06 for 10-year CVD prediction using GBT models with longitudinal inputs [128].

The range of findings between these studies can potentially be attributed to differences between study settings, risk factors included, and the frequency or quantity of repeated risk factor measurements. It has been suggested that predictive improvement, if any, from sequential risk factor measurements may reflect the ability to account for measurement error and variability in risk factors, with no substantial further benefit derived from estimating individual-level slopes of risk factors over time [124, 126, 127]. Consequently, it is not surprising that the predictive benefit of longitudinal risk factor data can vary greatly depending on the particular characteristics of a study and the underlying cohort. While the present study had appropriate statistical power to detect substantial improvements in AUC of 0.05-0.06 reported by Cho *et al.* [127] and Zhao *et al.* [128], greater statistical power would be required to detect the much more modest improvements of less than 0.01 reported by Paige *et al.* [126]. Since AUC is well-known to be insensitive for detecting small differences in discriminative ability between models [133], we have also characterised the models in this study using NRI and IDI. The validity of using these metrics has been criticised for their risk of false positive indications of predictive differences between models [134]. However, even for these overly sensitive metrics, there was little or no evidence for differences in predictive performance between each modelling approach.

The limitations of the present study include (i) lack of data on certain established risk factors for stroke such as atrial fibrillation (AF) and blood cholesterol levels; (ii) quality of self-reported questionnaire data about medical history of coronary heart disease or diabetes; and (iii) inclusion of only three sequential measurements of risk factors recorded over 10 years. Importantly, a recent population-based study of Chinese adults involving comparable age groups [106], reported that the prevalence of AF is low in China (0.71% in adults aged 35 years or older). Consequently, although AF is an important predictor of stroke, its omission is likely to have a material effect for only a small proportion of individuals in the present study.

Similarly, mean blood levels of total cholesterol and LDL cholesterol are well-known to be lower in China than in Western populations and may be less important for stroke risk prediction in this population [135]. To adjust for errors in self-reported prior medical history, we considered an individual to have a history of coronary heart disease or diabetes if they were ever reported or detected (e.g., by blood glucose screening) to have such disease on any previous visit. Finally, while the present study included only three sequential measurements of risk factors recorded over 10 years, increased number or frequency of measurements may still be useful for improving stroke prediction and should be investigated further.

6.5 Conclusion

Individuals at highest risk for stroke in China from 2013-2014 were largely the same individuals who had the fastest growing risk of stroke throughout the previous decade. However, in this population, measurements of CVD risk factors for prediction of stroke at three separate surveys had limited additional utility compared to those recorded at a single survey. These results provide further support for current guidelines for clinical practice, which suggest that stroke prediction is best estimated and cost-effectively implemented using models involving measurements of major risk factors recorded at a single visit.

7

Optimal Statin Therapy Strategies for Primary Stroke Prevention

7.1 Introduction

As emphasised throughout this thesis, risk prediction models for stroke and cardiovascular disease (CVD) are important tools for identifying high-risk individuals who may benefit from use of blood pressure-lowering and lipid-lowering medication. While Chapters 4 to 6 have focused on the development of novel stroke risk models for Chinese individuals, in this chapter, we turn our attention to the application of such risk prediction tools for informing personalised treatment decisions. In particular, we focus on the controversial use of lipid-lowering statin therapy in China.

Statin use has increased worldwide following dissemination of evidence for the beneficial effects of statins for primary and secondary prevention of CVD [136]. However, it is uncertain if the benefits outweigh the risks for all individuals with high risk of CVD, especially for those with high risk of haemorrhagic stroke (HS). This question is particularly relevant for low- and middle-income countries (LMICs) such as China, where incidence of HS is much higher than in high-income countries (HICs) [6].

Randomised trials have demonstrated that use of statins (3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitors) to lower low-density lipoprotein cholesterol

(LDL-C) levels is highly effective for primary and secondary prevention of CVD [136], and guidelines in most countries have advocated a wider use of statins for CVD prevention in high-risk individuals [102, 118, 137]. However, the prevalence of statin use is much lower in Chinese adults compared to those in HICs [138, 139]. Wider use of statins could improve primary prevention of CVD in China [20, 140, 141], but obstacles include cost, lower tolerance of statins in Chinese adults, and concerns about increasing risks of HS [20, 102, 140, 142].

While statins reduce risk of ischaemic stroke (IS) and major coronary events (MCEs: myocardial infarction and fatal ischaemic heart disease), they also increase the risk of HS [20, 136]. This increased risk is particularly relevant in China and other LMICs, where HS accounts for a larger proportion of stroke events than in HICs and HS also has a higher case fatality than IS [7]. Thus, while previous studies have reported the net benefit of prescribing statins to individuals with elevated CVD risk [20, 140, 141], substantial uncertainty persists about whether there are individuals with particularly high HS risk for whom statin treatment should be avoided. Most previous risk prediction models assessed risk of total stroke or total CVD as a single endpoint and did not distinguish major pathological stroke types [8, 10, 46, 47]. One study estimated 10-year risk prediction models for IS and HS in three population-based cohort studies, but did not include Chinese individuals and did not assess the models for informing personalised treatment decisions [11].

The aim of the study presented in this chapter was to fill this knowledge gap by developing novel risk models to identify individuals whose first stroke event was most likely to be haemorrhagic, and to use these models to optimise statin treatment decisions to minimise stroke events or deaths in a contemporary population of 0.5 million Chinese adults. We first considered primary prevention of stroke alone, and then extended our analyses to include both stroke and MCEs.

7.2 Methods

7.2.1 Study Design and Participants

Details of the design and methods used in the China Kadoorie Biobank (CKB) are reported in detail in Chapter 3. As in Chapters 4 and 5, the analyses presented in this chapter were restricted to CKB individuals with no prior history of stroke or transient ischaemic attack (TIA) at the date of recruitment (205,293 men, 298,549 women). The participants were followed-up to detect stroke, MCEs, and death until January 1, 2018. However, unlike in previous chapters, all incident cases of first stroke and MCEs were included, and were not limited to only those recorded during the first nine years of follow-up. In addition to coding of IS and HS cases using ICD-10 (described in Section 3.2), MCEs were also coded by trained medical staff, with further review done centrally by a clinical research fellow. MCEs were defined to include myocardial infarction (I21–I23) and fatal ischaemic heart disease (I20–I25).

7.2.2 Statistical Analysis

Among 503,842 individuals with no prior history of stroke or TIA at the date of recruitment, 49,987 first incident cases of IS, 11,813 first incident cases of HS, and 11,744 incident cases of MCEs were recorded by January 1, 2018, with only 1% of participants lost to follow-up. Baseline characteristics including demographic factors, lifestyle factors, physical measurements, medical history, and health status were compared between individuals whose first incident stroke was IS, individuals whose first incident stroke was HS, and those who had no incident stroke. We calculated the age-specific incidence of first-ever stroke and first stroke death for IS and HS stratified by sex, type of area (urban/rural), and vascular risk. Low vascular risk populations were defined as those with no measured hypertension or prior history of coronary heart disease (CHD). Medium-risk populations were defined as those with measured hypertension, but with no prior history of CHD. High-risk populations were defined as those with prior history of CHD. Hypertension was defined as measured systolic blood pressure (SBP) of ≥ 140 mmHg, or a measured

diastolic blood pressure (DBP) of ≥ 90 mmHg, or receiving drug treatment for hypertension. Since some individuals were recruited as late as 2008, incidence rates were calculated as nine-year risks using one minus Kaplan-Meier survival probability.

To identify individuals most likely to have a first stroke event that is HS (and thus be most likely to be harmed by statin treatment), we compared two modelling approaches. A visual overview of both approaches is provided in Figure 7.1. For both approaches, CKB individuals were randomly divided into a training set (80%) and test set (20%) for model development and internal validation, respectively, with the training set and test set being the same for both approaches.

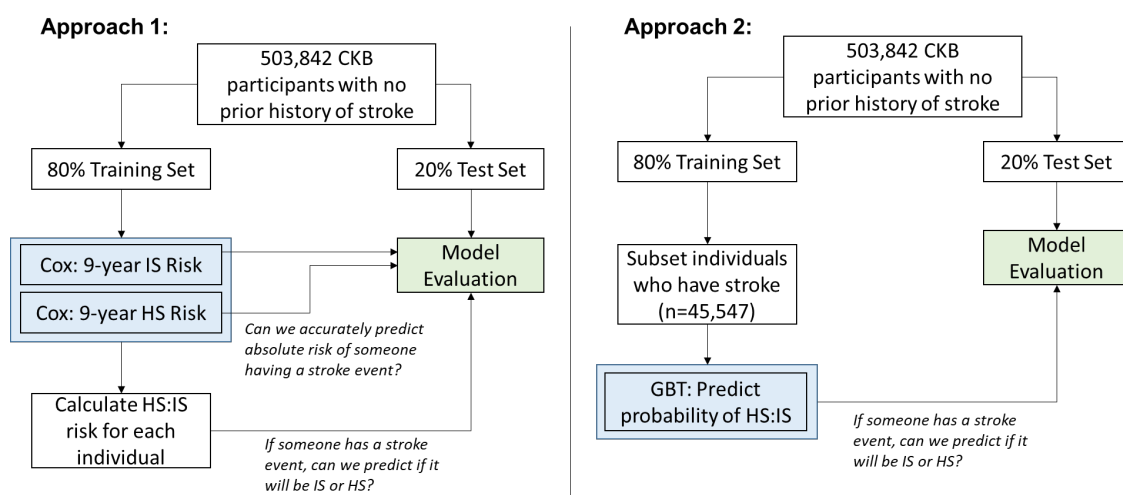


Figure 7.1: Overview of two modelling approaches for identifying individuals whose first stroke event is most likely to be haemorrhagic.

In the first modelling approach, we developed sex-specific and area-stratified Cox models to separately predict nine-year absolute risks of IS and HS, similar to traditional risk prediction scores for stroke and CVD [8, 10, 46, 47], and compared each individual's predicted ratio of HS risk-to-IS risk. Cox model development was performed using an identical methodology to the development of the “+ Additional risk factors” models described in Chapter 4 (Section 4.2.2). Cox models were individually evaluated for discrimination (using the AUC metric) and calibration (using calibration plots) in separately predicting nine-year risk of IS and HS. Then, after ranking individuals by predicted ratio of HS risk-to-IS risk, the overall approach was evaluated for its ability to discriminate individuals who had a first HS event

from individuals who had a first IS event using the AUC metric and by analysing rates of IS and HS by deciles of predicted HS risk-to-IS risk.

In the second modelling approach, we developed sex-specific gradient boosted tree (GBT) models that were trained only on individuals in the training set who experienced an incident stroke ($n=45,547$) and directly predicted the probability that an individual's first stroke would be HS conditional on a first stroke event occurring. Additional details of the GBT model development are provided in Appendix E.1. This modelling approach was also assessed for its ability to discriminate individuals who had a first HS event from individuals who had a first IS event and was compared to the first approach. The superior approach was used to assess optimal statin therapy policies to minimise stroke events and fatalities.

To estimate the effects of lowering LDL-C by 1 mmol l^{-1} with statins, we divided the CKB test set by age and by vascular risk, determining in each subgroup the optimal predicted HS risk-to-IS risk threshold below which statins should be prescribed to maximise: (i) the number of primary stroke events avoided and (ii) the number of first stroke deaths avoided. The absolute numbers of strokes and deaths avoided for each subgroup were calculated assuming that lowering LDL-C by 1 mmol l^{-1} reduces the risk of IS by 20% [136], and increases risk of ICH by 17% [20], for both fatal and non-fatal events. We then combined these subgroup estimates by standardising to the overall CKB population. To consider the effect of statins on MCEs, we performed a similar procedure by also including incident MCEs that occurred before any stroke event, assuming a 24% risk reduction from lowering LDL-C by 1 mmol l^{-1} with statins [136]. The 95% CIs were estimated by resampling 100 bootstrapped samples from the test set.

LASSO variable selection for the Cox models was performed in R version 3.6.1 using the glmnet package [86] version 3.0-2. All other statistical analyses were performed using Python version 3.7.0. Cox proportional hazards models were implemented using the lifelines package [95] version 0.21.1. GBT models were implemented using the scikit-learn toolkit [96] version 0.19.2.

7.3 Results

7.3.1 Baseline Characteristics

Among the 503,842 CKB study participants in the present analyses, the mean (SD) age was 51.9 (10.6) years and 59% were women. Among a total of 57,802 individuals who had an incident stroke event, 48,917 (84.6%) had a first IS and 9,808 (17.0%) had a first HS (Table 7.1). Compared with those who had no stroke, individuals who had a first stroke were older and more likely to have prior history of CHD, diabetes, and hypertension. Stroke cases had a lower income and were more likely to have poor self-rated health status. Compared to individuals whose first stroke was IS, individuals who had a first HS were less likely to be women (49.2% vs 56.7%) and less likely to live in urban areas (28.2% vs 54.8%). In addition, individuals with a first HS tended to have a lower income, be less educated, have lower body mass index (BMI), and report greater physical activity than individuals with a first IS. They also had a lower prevalence of diabetes (8.2% vs 12.7%) and CHD (4.6% vs 8.0%), included a slightly higher proportion of ever-regular smokers (75.8% vs 72.8%) and had substantially higher prevalence of hypertension (67.6% vs 56.1%) than those who had a first IS. Overall, men and women had similar rates of prior history of CHD (2.5% vs 3.0%), diabetes (5.3% vs 6.0%), and use of blood pressure-lowering medication (9.9% vs 11.4%), but ever-regular smoking and drinking was much more common in men than in women, and were consequently stratified by sex in Table 7.1.

7.3.2 Incidence of Primary Ischaemic Stroke and Haemorrhagic Stroke

Nine-year incidence of both IS and HS increased rapidly with age in both sexes and in urban and rural areas of China (Figure 7.2). The overall incidence of IS was much higher than the incidence of HS, and the incidence of both stroke types were higher in men than in women, with more extreme sex differences for HS than for IS. Incidence of IS was higher in urban than in rural areas, but the incidence of HS was higher in rural than in urban areas.

Table 7.1: Baseline characteristics of included participants from the China Kadoorie Biobank, stratified by pathological type of first stroke event.

	First IS	First HS	No Stroke
Number of participants	48,917	9,808	446,040
<i>Demographic factors</i>			
Age at baseline, mean (s.d.) (years)	59.4 (9.6)	59.3 (10.2)	50.9 (10.4)
Age 30-39 years, (%)	2.6	4.1	17.0
Age 40-49 years, (%)	15.1	15.3	32.1
Age 50-59 years, (%)	32.7	30.2	30.6
Age 60-69 years, (%)	33.8	32.9	15.4
Age 70-79 years, (%)	15.9	17.5	4.9
Female (%)	56.7	49.2	59.7
Urban (%)	54.8	28.2	43.0
≥6 years of education (%)	46.3	30.9	49.9
Household income (≥20,000 yuan per year) (%)	37.5	29.0	43.7
<i>Medical history and health status</i>			
Diabetes (%)	12.7	8.2	4.9
Coronary heart disease (%)	8.0	4.6	2.2
Hypertension (%)	56.1	67.6	30.7
Blood pressure drugs (%)	22.6	23.0	9.2
Self-rated poor health status (%)	13.4	15.6	9.4
<i>Physical measurements, mean (s.d.)</i>			
SBP (mmHg)	140.9 (24.0)	149.4 (26.3)	129.3 (20.2)
DBP (mmHg)	80.7 (12.3)	84.8 (13.9)	77 (10.8)
BMI (kg m ⁻²)	24.4 (3.5)	23.4 (3.6)	23.6 (3.3)
<i>Lifestyle factors</i>			
Male ever-regular smokers (%)	72.8	75.8	74.5
Female ever-regular smokers (%)	5.8	6.0	2.9
Male ever-regular drinkers (%)	42.4	41.7	41.9
Female ever-regular drinkers (%)	3.2	3.7	2.9
Physical activity, mean (s.d.) (hours per day)	4.9 (3.1)	5.3 (3.3)	6.6 (3.4)

Ever-regular smokers include current- and ex-regular smokers.

Ever-regular drinkers include current-regular, ex-regular and reduced-intake alcohol drinkers.

923 individuals had a first IS and first HS recorded on the same date and were included in both columns.

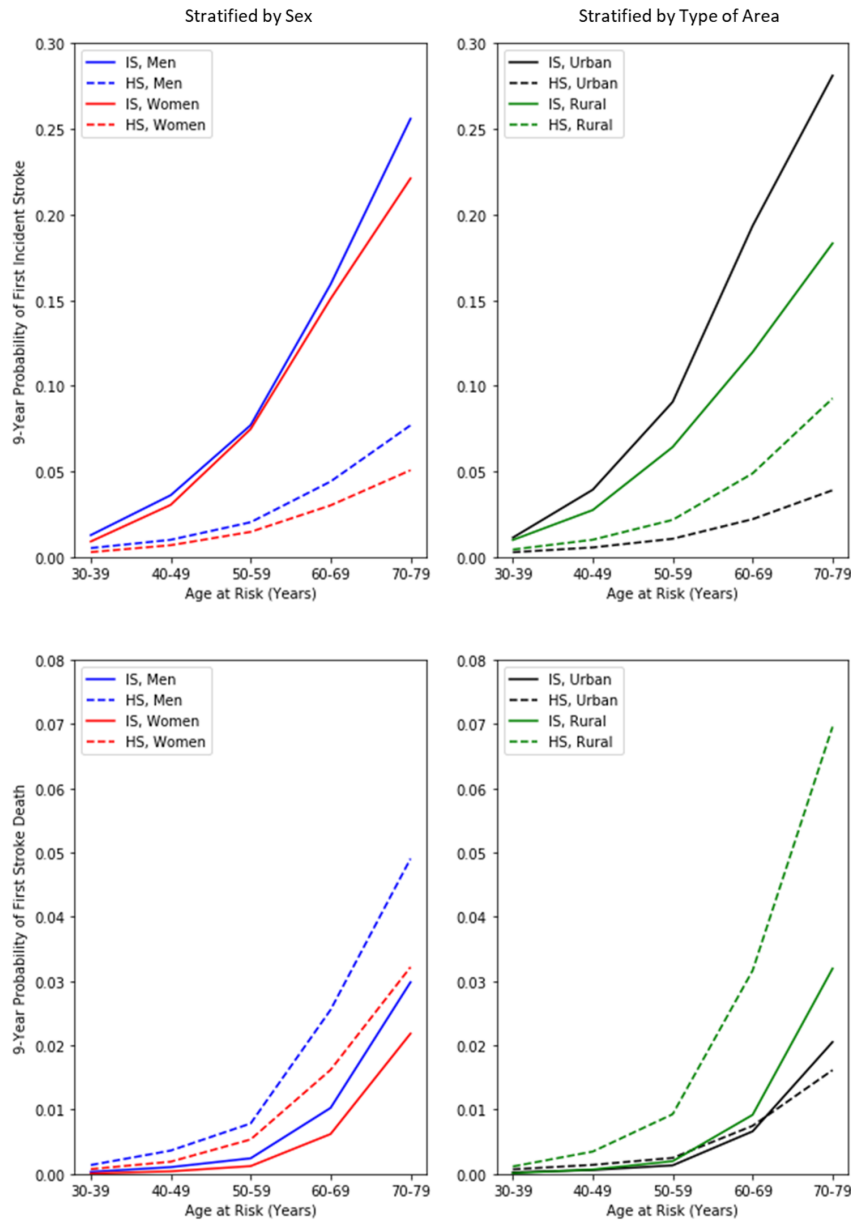


Figure 7.2: 9-year probability of first incident stroke and first stroke death from first ischaemic stroke (IS) and first haemorrhagic stroke (HS) cases by age group with stratification by sex (left) and type of area (right).

Mortality caused by first incident strokes also increased greatly with increasing age for both IS and HS. Importantly though, first incident HS events had much higher case fatality than first incident IS, and incident mortality from first HS events was actually higher than that from IS in both sexes despite HS having a substantially lower incidence (Figure 7.2). Mortality for both IS and HS was higher

in rural than in urban areas, with mortality from IS in rural areas actually exceeding mortality from HS in urban areas for individuals aged ≥ 60 years.

Analysis of incidence and mortality of IS and HS by vascular risk groups yielded similar findings (Appendix: Figure E.1). However, while incidence of IS increased with vascular risk, incidence of HS was higher for the medium-risk population than the high-risk population, most likely due to hypertension being more strongly associated with HS than prior history of CHD (Table 7.1).

7.3.3 Discrimination of Ischaemic Versus Haemorrhagic First Stroke Cases

Similar to the findings of Chapter 4, the sex-specific Cox models for separately predicting absolute risks of each major stroke type independently yielded good discrimination for both IS (AUC [95% CI]: 0.828 [0.821–0.833] in men; 0.831 [0.825–0.837] in women) and HS (AUC: 0.815 [0.800–0.829] in men; 0.826 [0.812–0.838] in women) in addition to good calibration (Appendix: Figures E.2–E.3). The leading risk factors (ordered by effect size) for IS and HS included age, blood pressure-related variables, diabetes, smoking, alcohol consumption, health insurance coverage, and family medical history (Tables 7.2–7.3). Plotting predicted HS risk against IS risk for each individual demonstrated different patterns by sex and area, and further demonstrated that only a small proportion of individuals had higher absolute risk for HS than for IS (and these were almost exclusively from rural areas) (Figure 7.3). Moreover, ranking individuals by ratio of predicted HS risk-to-IS risk was effective for discriminating the individuals whose first stroke event was most likely to be haemorrhagic (AUC: 0.731 [0.719–0.743]), although even for individuals in the highest decile of HS risk-to-IS risk, absolute rates of IS were still higher than those of HS (Figure 7.4).

The GBT modelling approach also yielded good discrimination (AUC: 0.744 [0.733–0.756]), slightly outperforming the Cox modelling approach for identifying individuals who would have HS (Figure 7.4). However, unlike the Cox models, which selected features for their utility in independently predicting IS and HS, the

Table 7.2: Top 10 important features by effect size in Cox models for ischaemic stroke (IS) in men and women.

Feature Ranking	Men		Women	
	Feature Name	Hazard Ratio [95% CI]	Feature Name	Hazard Ratio [95% CI]
1	Age (10 years)	1.82 [1.78-1.88]	Age (10 years)	1.90 [1.84-1.95]
2	Used BP drugs	1.59 [1.47-1.71]	Smokes most days	1.74 [1.16-2.59]
3	Diabetes	1.30 [1.20-1.40]	Used BP drugs	1.38 [1.30-1.47]
4	Smokes daily	1.28 [1.24-1.32]	Rheumatic heart disease	1.36 [1.11-1.67]
5	Diabetes, if age <65	1.28 [1.16-1.40]	Diabetes	1.32 [1.24-1.40]
6	Missing mother medical history	0.79 [0.51-1.23]	Health coverage	1.26 [1.22-1.31]
7	Health coverage	1.25 [1.19-1.31]	Age 65	0.81 [0.78-0.85]
8	Ex-regular drinker <3 bowel movements per week	1.24 [1.15-1.34]	Children stroke	1.22 [0.93-1.59]
9	Hypertension diagnosis	1.23 [1.12-1.36]	Hypertension diagnosis	1.21 [1.16-1.27]
10			Worse comparative health	1.21 [1.16-1.26]

Table 7.3: Top 10 important features by effect size in Cox models for haemorrhagic stroke (HS) in men and women.

Feature Ranking	Men		Women	
	Feature Name	Hazard Ratio [95% CI]	Feature Name	Hazard Ratio [95% CI]
1	Used BP drugs	1.86 [1.61-2.14]	Used BP drugs	2.05 [1.80-2.34]
2	Age (10 years)	1.49 [1.41-1.58]	Rheumatic heart disease	1.64 [1.12-2.40]
3	House wife/husband	1.36 [1.18-1.55]	Age (10 years)	1.56 [1.47-1.66]
4	Diabetes	1.33 [1.20-1.48]	Ex-regular drinker	1.31 [0.98-1.73]
5	Worse comparative health	1.27 [1.15-1.40]	Worse comparative health	1.26 [1.15-1.38]
6	Father cancer	0.81 [0.72-0.92]	Never smoker	0.82 [0.73-0.92]
7	Mother still alive	1.21 [1.11-1.32]	Diabetes	1.22 [1.10-1.34]
8	Ex-regular drinker	1.21 [1.08-1.36]	Father still alive	1.15 [1.05-1.27]
9	Hypertension diagnosis	1.20 [1.10-1.31]	Mother still alive	1.15 [1.06-1.25]
10	Eats rice monthly	1.19 [1.04-1.35]	Siblings stroke	1.14 [1.04-1.25]

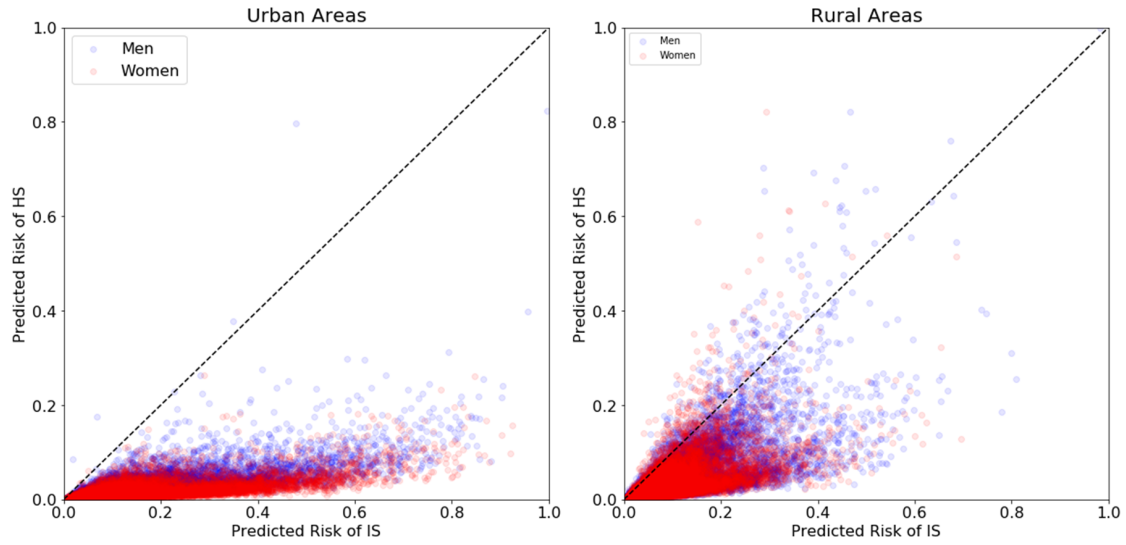


Figure 7.3: Cox risk estimates for 9-year risk of IS and HS in urban areas and rural areas. Red points indicate women. Blue points indicate men. All individuals represented were from the test set.

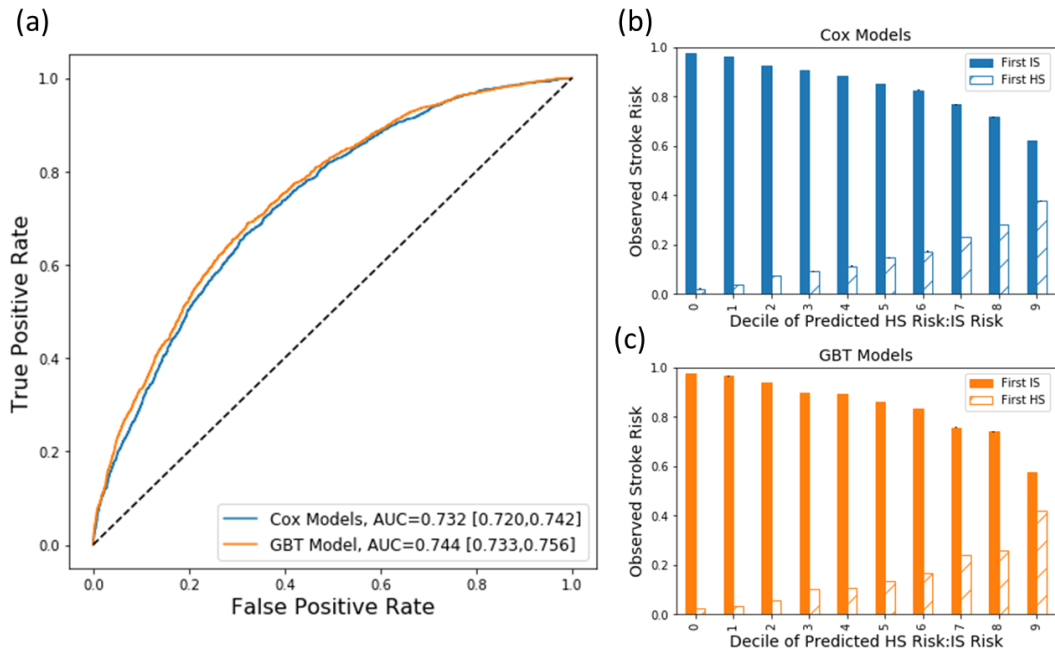


Figure 7.4: Comparison of discrimination of individuals with first IS and first HS between Cox and GBT models. Subplot (a) shows the receiver-operating characteristic curves and AUCs for both modelling approaches. Subplots (b) and (c) show the rates of IS and HS by deciles of HS risk-to-IS risk as predicted by the Cox modelling approach and GBT modelling approach, respectively.

Table 7.4: Top 20 important features used by GBT models for discriminating between a first ischaemic stroke (IS) and first haemorrhagic stroke (HS) event.

Feature Ranking	Men		Women	
	Feature Name	Gini Importance	Feature Name	Gini Importance
1	Urban region	0.081	Urban region	0.095
2	DBP	0.067	DBP	0.085
3	Age	0.058	Fat percent	0.061
4	Weight	0.055	Harbin	0.058
5	Fat percent	0.052	Age	0.057
6	SBP	0.050	SBP	0.048
7	Harbin	0.048	Retired	0.048
8	Hunan	0.047	Gansu	0.047
9	Sitting height	0.035	Weight	0.033
10	SBP, no HTN Rx	0.031	SBP, no HTN Rx	0.030
11	Physical activity (MET-h per day)	0.029	Physical activity (MET-h per day)	0.030
12	Henan	0.029	Hunan	0.027
13	Standing height	0.027	Sitting height	0.025
14	BMI	0.025	Standing height	0.025
15	Waist size	0.024	BMI	0.022
16	Physical activity (hours per day)	0.023	Physical activity (hours per day)	0.021
17	Gansu	0.022	Waist size	0.021
18	Heart rate	0.022	Haikou	0.020
19	Daily consumption of staple foods other than rice/wheat	0.022	Henan	0.020
20	Retired	0.021	Daily consumption of staple foods other than rice/wheat	0.017

Gini importance is computed as the normalised total reduction of the Gini impurity brought by each feature. Higher Gini importance values correspond to more important features.

GBT models specifically selected features for their utility in distinguishing between the major stroke types, including age, blood pressure-related variables, regional information, weight, and body fat percentage (Table 7.4). Moreover, compared to the Cox modelling approach, the individuals identified by the GBT models to be in the highest decile of HS risk-to-IS risk had slightly higher rates of HS and lower rates of IS, suggesting better identification of individuals who may have the highest risk of harm from statin treatment. For these reasons, the GBT modelling approach was selected as the superior approach for further investigating the optimal use of statin therapy in Chinese adults.

7.3.4 Optimal Use of Statin Therapy to Prevent Stroke Events and Stroke Deaths

Driven by the higher incidence of IS than HS, the optimal use of statins in CKB to minimise stroke events would advocate treatment of 98.9% [96.2%–99.0%] of CKB individuals, excluding only a small number of individuals with the highest predicted HS risk-to-IS risk ratio (Figure 7.5). Such a policy could have avoided 118 [115–122] events per 10,000 CKB individuals during the follow-up period, but would result in an excess of 4 [3–5] deaths per 10,000 individuals due to an increased risk of HS.

In contrast, the optimal use of statins to minimise stroke deaths would advocate treatment of a smaller proportion (31.7% [15.1%–41.7%]) of CKB individuals, accounting for the higher case fatality of HS compared to IS (Figure 7.5). Such a policy could still avoid 61 [37–71] stroke events per 10,000 CKB individuals during the follow-up period, while also avoiding 0.4 [0.3–0.7] deaths. For both policies (i.e., minimising stroke events and minimising stroke deaths), the estimated effects of statins were generally greater in older individuals with higher vascular risk.

7.3.5 Additional Inclusion of Major Coronary Events

Among previously stroke-free individuals, incidence of MCEs was comparable with that for HS, but had a higher case fatality rate (53.3% vs 31.4%) (Table 7.5). Age-specific incidences of MCEs stratified by sex, area type, and vascular risk are included in Appendix: Figure E.4 as well as baseline characteristics of individuals who had an incident MCE prior to any stroke event (Appendix: Table E.1). Consequently, upon considering the effects of statins on reducing risk of MCEs as well as IS, the optimal statin treatment policies were very similar both for minimising events and for minimising deaths. Both policies would advocate treatment of nearly all individuals, with the exception only of those with the very highest predicted HS risk-to-IS risk ratios (Figure 7.6). To minimise events, the optimal proportion of CKB treated was 99.6% [97.7%–99.7%], which was estimated to avoid 149 [146–153] events and 12 [10–13] deaths per 10,000 individuals. Meanwhile, to minimise deaths, the optimal

proportion of CKB treated was 97.5% [86.9%–98.3%] of CKB individuals, and was estimated to avoid 147 [141–151] events and 12 [11–14] deaths per 10,000 individuals.

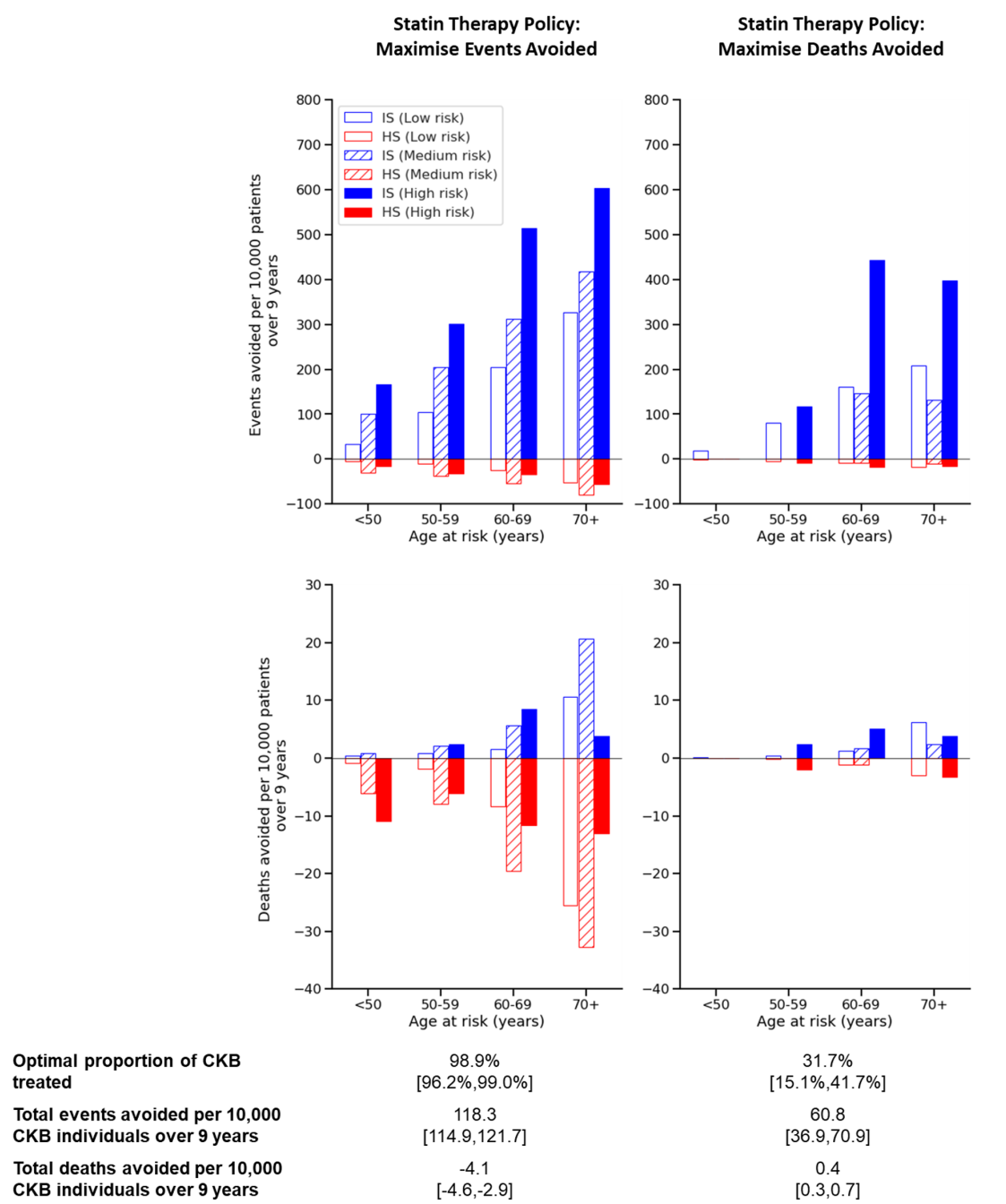
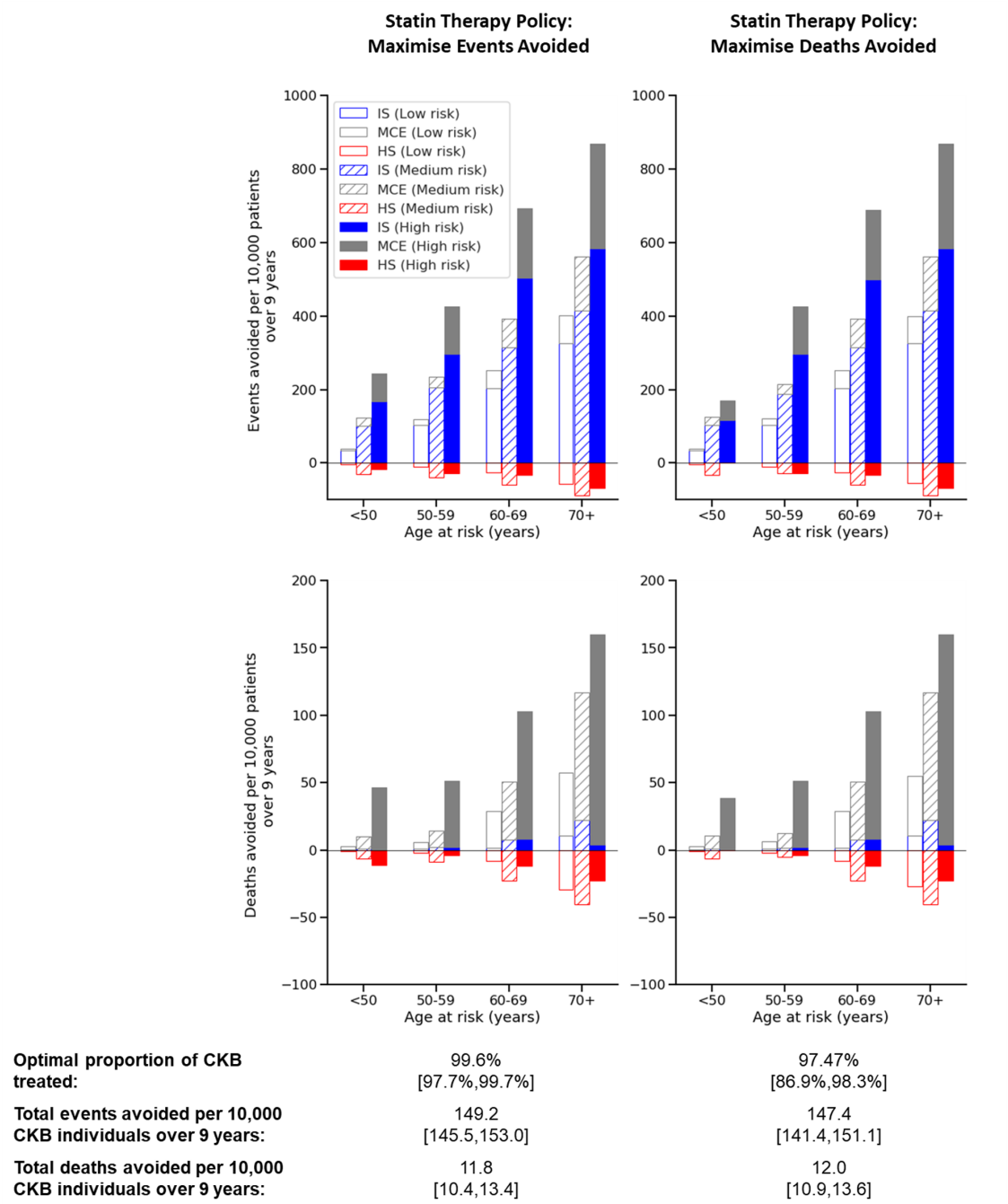


Table 7.5: Numbers and case fatality of reported first events of ischaemic stroke (IS), haemorrhagic stroke (HS), and major coronary events (MCE).

First Event Type	Individuals with Event (% Fatal)						
	Men (n=205,293)	Women (n=298,549)	Urban (n=220,924)	Rural (n=282,918)	Low Risk (n=327,631)	Med Risk (n=161,966)	High Risk (n=14,245)
No IS/HS/MCE	174,880 (NA)	262,737 (NA)	188,199 (NA)	249,418 (NA)	301,681 (NA)	127,087 (NA)	8,849 (NA)
IS	20,872 (2.3%)	27,505 (1.5%)	26,446 (1.5%)	21,931 (2.4%)	19,918 (1.5%)	24,641 (2.2%)	3,818 (1.6%)
HS	4,936 (33.0%)	4,797 (29.8%)	2,734 (16.3%)	6,999 (37.3%)	3,053 (30.5%)	6,245 (31.6%)	435 (34.9%)
ICH	4,663 (34.7%)	4,331 (32.4%)	2,441 (17.5%)	6,553 (39.6%)	2,682 (34.0%)	5,897 (33.2%)	415 (35.9%)
SAH	356 (5.6%)	587 (6.0%)	390 (7.4%)	553 (4.7%)	464 (5.0%)	448 (5.8%)	31 (19.4%)
MCE	5,261 (52.1%)	4,217 (54.9%)	4,260 (42.4%)	5,218 (62.3%)	3,402 (51.1%)	4,820 (55.4%)	1,256 (51.9%)

ICH: intracerebral haemorrhage. SAH: Subarachnoid haemorrhage.
A small proportion (0.3%) of individuals had multiple event types recorded on the same date and are included in multiple corresponding rows.



for primary prevention of stroke and MCEs. Such concerns have been an obstacle to the wider use of statins in China, with <5% of CKB individuals at high risk of CVD reporting regular use of statins at enrolment [20], consistent with other Chinese studies [138, 139]. Although the incidence of HS is higher in China than in HICs (where statin usage in high-risk individuals is approximately 66% [139]), we demonstrated that in over 97% of Chinese adults, the benefits of statins still outweigh the potential hazards due to the substantially higher absolute rates of IS compared to HS. While an argument can be made for treating a smaller proportion of the Chinese population in an effort to prevent excess stroke deaths (i.e., due to the higher case fatality of HS), such benefits are outweighed after considering the risk reduction effect of statins on MCEs, which are equally prevalent and even more fatal than HS in previously stroke-free Chinese adults. Moreover, the benefits of statin treatment in China would be even more pronounced if current trends of increasing IS incidence and decreasing HS incidence continues [143, 144], or if the beneficial effect of statins on other occlusive vascular diseases are also considered [136, 145]. Furthermore, while we only provide estimates of avoidable events and deaths from primary stroke events and MCEs, the net benefits of statins are likely to be even greater due to downstream prevention of secondary CVD events and associated deaths.

Previous studies have reported the benefits of wider statin treatment in China for prevention of CVD, but have either failed to separately consider the increased risk of HS [140, 141], or simply argued that the overall risks of excess HS events are outweighed by the benefits of statins for reducing IS and MCEs at a population level [20]. However, it was unknown if there are individuals with especially high risk of HS for whom statin therapy should be contraindicated. Few previous studies have explored the prediction of HS risk, and these have been limited by small case numbers and/or lack of representation from Chinese individuals [11, 146, 147]. For example, one study developed separate 10-year risk scores for IS and HS in individuals from the Atherosclerosis Risk in Communities Study, the Rotterdam Study, and the Cardiovascular Health Study, but included only a few hundred HS

cases among 28,000 individuals, explicitly excluded Chinese participants, and did not explore the use of their models for personalised treatment decisions [11].

In contrast, in this study, we developed two separate modelling approaches that each reliably identified individuals whose first stroke event was most likely to be haemorrhagic, and we demonstrated that even among the majority of these individuals, the benefits of statins substantially exceeded any potential hazards. Moreover, by uniquely considering each individual's differential risk for IS and HS, we provided a novel approach for identifying the small proportion of individuals (< 3%) for whom hazards for statin therapy do appear to exceed the benefits.

This study had several strengths, including the prospective study design with very low loss to follow-up, the large numbers of Chinese adults, and the high numbers of well characterised stroke events (92% confirmed by brain imaging). However, the study also had several limitations. Firstly, while the CKB included individuals from 10 geographically diverse areas in China, it was not designed to be representative of and may not be directly generalisable to the overall Chinese population. Secondly, recorded stroke events were limited to death from stroke and reported hospitalised strokes. Thirdly, the CKB lacked data on some important stroke risk factors including atrial fibrillation (AF) and blood levels of total cholesterol, which could have enhanced model performance to identify individuals at higher risk of suffering a first HS event. However, a recent population-based study of Chinese adults involving comparable age groups [106], reported that the prevalence of AF is low in Chinese adults (0.71% in adults aged 35 years or older). Consequently, although AF is an important predictor of stroke, its omission is unlikely to have a material effect as only a small proportion of individuals in the present study are affected. Similarly, mean blood levels of total cholesterol and LDL cholesterol are lower in Chinese than in Western populations and may be less important for prediction of stroke risks in this population [135]. Finally, while the findings of the present study suggest that nearly all Chinese adults could benefit from statin therapy, implementing treatment decisions in practice should also incorporate economic considerations such as the cost of reducing events and deaths in different age and risk groups as well

as the expected number of quality-adjusted life years (QALYs) gained from such policies. For example, while focussing treatment on high-risk patients aged ≥ 70 years could avoid 801 events and 137 deaths per 10,000 individuals, targeting low-risk patients aged < 50 years would only prevent 35 events and 2 deaths per 10,000 individuals (Figure 7.6). On the other hand, the QALYs gained by preventing each event is likely to be higher in the younger, low-risk adults. Current international guidelines recommend statin treatment for individuals who exceed a 10-year CVD risk threshold (e.g., 7.5% in the United States, 10% in the United Kingdom), and similar policies may be considered in China [118, 119].

7.5 Conclusion

In conclusion, using novel modelling approaches for assessing the differential risk of IS and HS in Chinese individuals, we have demonstrated that the potential hazards of statins from increasing HS risk are substantially exceeded by the benefits from reducing risk for IS and MCEs, even in the majority of individuals at highest risk of HS. The findings address an important issue that has been a major obstacle to the wider use of statins in Chinese adults, and provide support for wider use of statins, as advocated by both recent Chinese and international guidelines [102, 118, 137].

8

Heterogeneity and Classification of Ischaemic Stroke Subtypes

8.1 Introduction

In this chapter, we extend our investigations beyond total stroke and its major pathological types (i.e., ischaemic stroke [IS] and haemorrhagic stroke [HS]), focusing on aetiological subtypes of IS in China.

China has more IS cases than any other country in the world, including 24 million prevalent IS cases and 1 million deaths in 2019 [1, 2]. As discussed in Chapter 2, stroke is a heterogeneous disease, and aetiological subtypes of IS (e.g., cardioaortic embolism [CE], large artery atherosclerosis [LAA], small artery occlusion/lacunar stroke [SAO]) are routinely used to classify IS cases based on their distinct pathophysiology and prognosis. Validated algorithms for classification of IS by aetiological subtypes include the Trial of Org 10172 in Acute Stroke Treatment (TOAST) [13] and Causative Classification System for Ischemic Stroke (CCS) [14, 24], which both classify IS cases using presenting symptoms, clinical signs, and investigations. However, in settings with incomplete investigations, conventional algorithms frequently fail to identify a single likely aetiology, leaving a large proportion of IS cases unclassified. Previous studies of IS subtypes in China have been constrained by studies involving a small number of cases or studies

involving a limited geographic area within China [27, 148–150]. Nevertheless, there is evidence that the distribution of IS subtypes and their associated risk factors differs between Chinese and European populations [27, 148, 149], and may have changed within China in recent decades [150].

We evaluated the distribution, risk factors, and stroke recurrence and mortality of IS subtypes in a contemporary population of over 0.5 million Chinese adults living in 10 diverse regions within China. Furthermore, to resolve incompletely investigated IS cases of undetermined aetiology, we developed a novel machine learning (ML) approach for probabilistically classifying IS cases using risk factor measurements recorded at enrolment in the study.

8.2 Methods

8.2.1 Study Design

Details of the design and methods used in the China Kadoorie Biobank (CKB) are reported in detail in Chapter 3. In addition to the standard coding of all fatal and non-fatal cases of stroke by trained medical staff using the International Classification of Diseases 10th Revision (ICD-10), all hospital-reported cases of first stroke also underwent separate clinical adjudication, which yielded additional data used in the present analyses. For these hospital-reported cases of first stroke, original medical records and brain imaging reports were retrieved and reviewed by clinical specialists in China. Adjudicators were asked to complete a standardised electronic form, which requested specific details on presenting symptoms and signs and diagnostic criteria, including imaging evidence of haemorrhage or ischaemia, laterality and anatomical location of any cerebral lesions, and additional evidence of extracranial and intracranial arterial disease or sources of embolism for classification by TOAST [13] and CCS [14, 24]. Based on the available information, expert stroke physicians were asked to adjudicate diagnosis of stroke types as either I63: IS (LACI [lacunar stroke], Non-LACI), I61: intracerebral haemorrhage, I60: subarachnoid haemorrhage, or I64: unspecified stroke types. Primary cases of

IS (n=22,216) were then classified into aetiological subtypes using an adapted version of CCS (Appendix F.1).

Table 8.1: Key variables collected from medical records for classification of ischaemic stroke (IS) subtypes using CCS.

Variable Name	Value Definitions
imaging_findings	- No brain imaging - Normal imaging - “Ischaemic (non-lacunar)”: >1.5 cm infarct - “Ischaemic (lacunar)”: ≤ 1.5 cm infarct. Irregular shape and undefined size (e.g., lacunae) - Indeterminate imaging
intra_extra_test_done	- No vascular imaging - Intracranial (IC) vascular imaging only - Extracranial (EC) vascular imaging only - Both IC and EC vascular imaging
intra_extra_is_abnormal	- No stenosis in either IC or EC artery - Any stenosis in IC artery - Any stenosis in EC artery - Any stenosis in both IC and EC artery
intra_extra_loc_art_dgre_stsis	- No stenosis $\geq 50\%$ in either IC or EC artery - Any stenosis $\geq 50\%$ in IC artery - Any stenosis $\geq 50\%$ in EC artery - Any stenosis $\geq 50\%$ in both IC and EC artery
ultrasound_plaq_featre	- NULL: No plaque - 0: If plaque, stable (hyperechoic) - EC artery only - 1: If plaque, unstable (hypoechoic or mixed echoic) - EC artery only
echo_ecg_test_done	- No cardiac test done - Electrocardiogram (ECG) test only - Echocardiography (Echo) only - Both ECG and Echo
ce_risk	- 0: No cardioembolic source - 1: Low- or uncertain risk cardioembolic source (e.g., mitral annular calcification, patent foramen ovale) - 2: High-risk cardioembolic source (e.g., atrial fibrillation, mechanical prosthetic valve)
systemic_embolization	- 0: No evidence of systemic embolism - 1: Evidence of systemic embolism
lacunar_syndrome	- 0: No clinical lacunar syndrome - 1: Presence of clinical lacunar syndrome (e.g., pure motor, pure sensory, ataxic hemiparesis, clumsy hand)

8.2.2 Data Used for CCS Classification of IS Subtypes

The key variables collected by the adjudicators for classification of IS subtypes using CCS are listed and defined in Table 8.1. The coverage of investigations such as

brain imaging, cardiac tests, and vascular tests is described in further detail below. However, to provide an initial visualisation of the data, we performed t-Distributed Stochastic Neighbor Embedding (t-SNE) using the key variables associated with these investigations. The resulting t-SNE plot (Figure 8.1) showed meaningful clustering by CCS-classified aetiology and provided useful insight into the CCS mapping of clinical and diagnostic findings to IS subtypes. For example, most IS cases with lacunar imaging findings were classified by CCS as SAO, most IS cases with potential cardiac sources of embolism were classified as CE, and many IS cases with abnormal vascular tests (i.e., stenosis in IC and/or EC artery) were classified as LAA. In general, strokes with fewer investigations were more likely to be classified by CCS as having undetermined aetiology, and overall, 30% of IS cases were undetermined. This highlights the shortcomings of CCS in settings with incomplete stroke investigation, in addition to the substantial gains that can be achieved from reliably and probabilistically classifying strokes of undetermined aetiology in CKB.

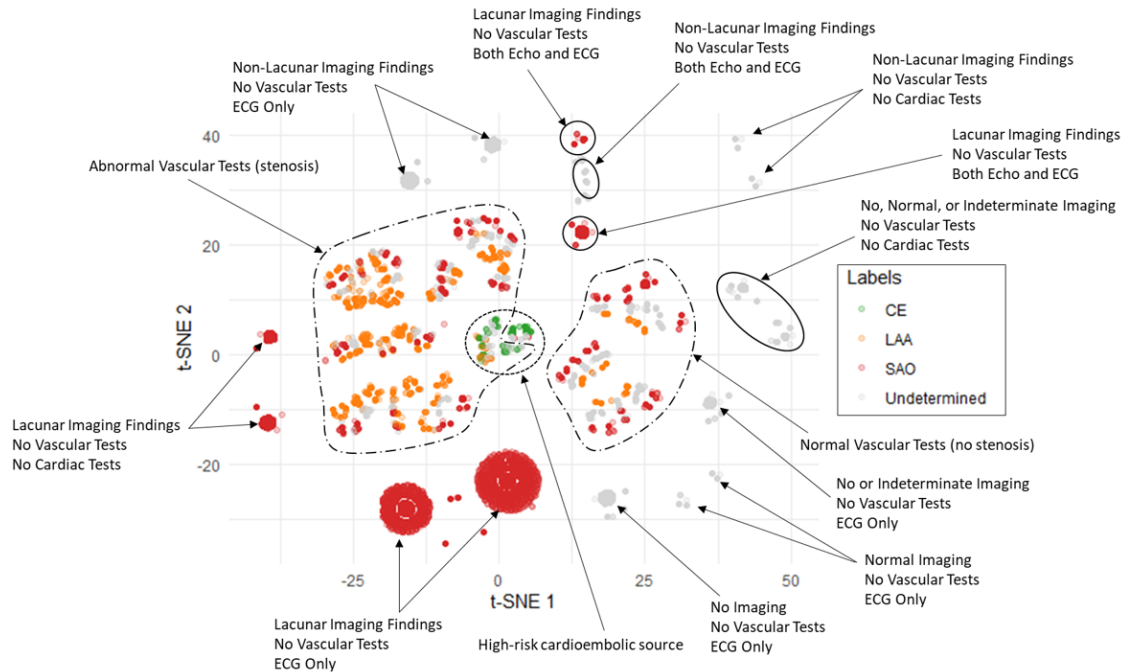


Figure 8.1: t-Distributed Stochastic Neighbor Embedding (t-SNE) visualisations of primary ischaemic stroke (IS) cases in the China Kadoorie Biboank coloured by CCS classification. CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion.

8.2.3 Statistical Analysis

The analyses presented in this chapter were restricted to individuals with no prior history of stroke or transient ischaemic attack (TIA) at the date of recruitment (205,293 men, 298,549 women). Baseline characteristics including demographic factors, medical history, health status, physical measurements, and lifestyle factors were compared between individuals whose first stroke event was CE, LAA, SAO, and undetermined (with incomplete investigation), as well as those who had no incident IS during the follow up period. Regional differences in overall IS incidence and distributions of IS subtypes were also evaluated. The cumulative all-cause mortality and recurrent stroke rates (measured as one minus Kaplan-Meier survival probability) following a first stroke were compared for each IS subtype. To maximise the validity of our findings, we considered “true” cases of each IS subtype to include cases of evident and probable aetiology classified by CCS. Recurrent strokes were defined as any hospitalised stroke events occurring more than 28 days after the onset of a known first stroke.

After investigating the differences between IS subtypes of determined aetiology, we randomly divided evident and probable cases of CE, LAA, and SAO into a training set (85%) and test set (15%) and developed a multiclass logistic regression (LR) model to classify IS cases based on baseline CKB risk factor variables, in addition to findings of cardioaortic sources of embolism [14, 24] identified from cardiac test results extracted from hospital records (Appendix F.2). Although not recorded at baseline in CKB, potential cardiac sources of embolism were included as model inputs because they are important for distinguishing CE from other IS subtypes, the determinants were present prior to the onset of stroke, and they can be cost-effectively detected with echocardiography (Echo) or electrocardiogram (ECG) tests. The LR model was evaluated on IS cases with determined aetiology in the test set using area under the receiver-operating characteristic curve (AUC), positive predictive value (PPV), sensitivity, F1 score, and accuracy. The model was also applied to incompletely investigated IS cases with undetermined aetiology, and evaluated by comparing outcomes of ML-classified cases of each IS subtype with

those of “true” cases with determined aetiology (an established method of validating subtype endpoints in the absence of ground truth) [151]. Incompletely investigated cases were defined as strokes with missing brain imaging (neither computerised tomography [CT] nor magnetic resonance imaging [MRI]), cardiac tests (neither Echo nor ECG), or vascular tests (neither intracranial nor extracranial vascular imaging).

Additional sensitivity analyses were performed with the removal of silent cerebral infarcts (i.e., imaging-detected cerebral infarctions without acute focal dysfunction), which are included in the revised ICD-11, but not the ICD-10 definition of stroke [28, 152]. All statistical analyses were performed using Python version 3.7.0. Kaplan-Meier estimates were implemented using the lifelines package [95] version 0.21.1, and the multiclass LR model was implemented using the scikit-learn toolkit [96] version 0.19.2.

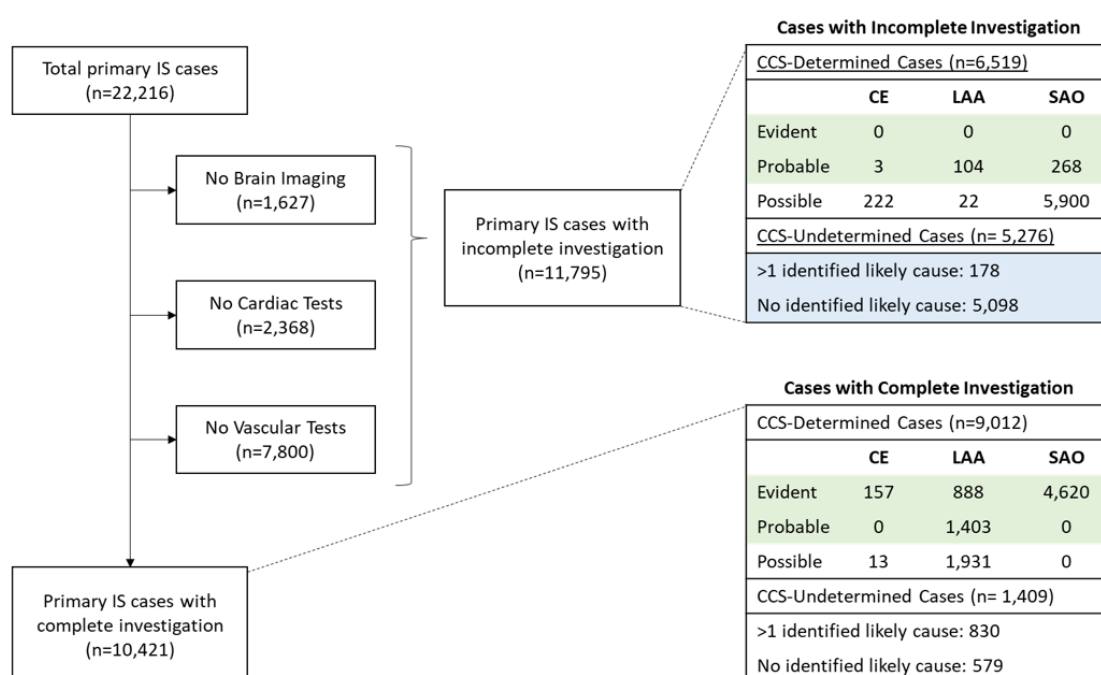


Figure 8.2: Flow chart of investigations and adjudicated subtypes of ischaemic stroke (IS) cases in the China Kadoorie Biobank. CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion. Green cells indicate evident and probable cases included in the analyses of each IS subtype. Blue cells indicated incompletely investigated cases with undetermined aetiology, classified in the present study using a novel machine learning approach.

8.3 Results

8.3.1 Heterogeneity of Ischaemic Stroke Subtypes

Of 22,216 primary IS cases, 70% had a determined causative subtype with 48% being classified as having evident and probable aetiology (Figure 8.2). 91% of IS cases had brain imaging, 87% had at least one cardiac test (Echo or ECG), and 51% had at least one vascular test (IC or EC vascular imaging), resulting in 11,795 cases with incomplete investigation. Importantly, 5,276 incompletely investigated cases had undetermined aetiology, representing 24% of all primary IS cases that may benefit from a probabilistic risk factor-based classification of IS subtypes when CCS lacks sufficient investigations to make a classification.

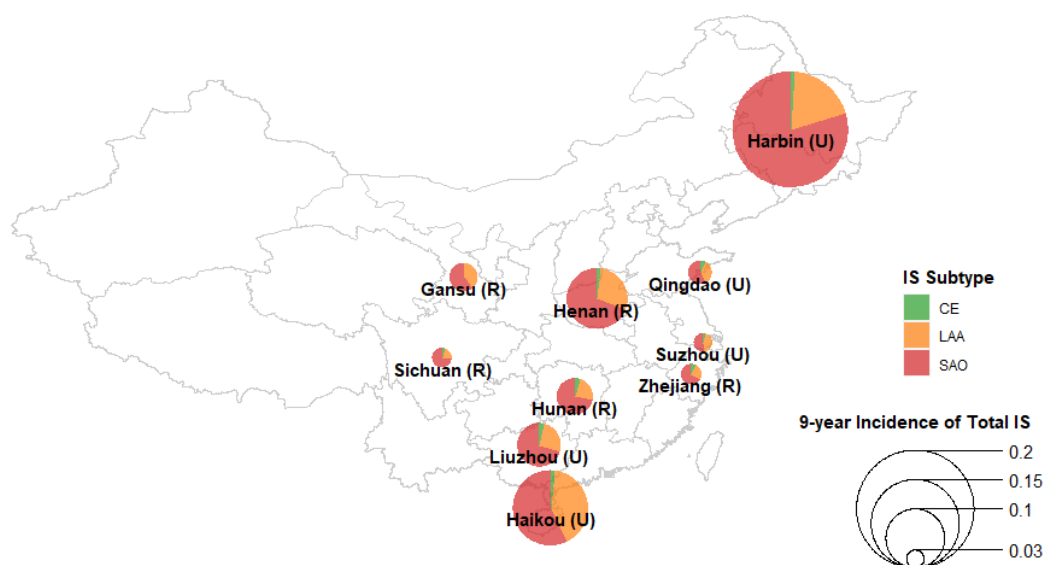


Figure 8.3: Proportions of ischaemic stroke (IS) subtypes by area in China. CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion. Proportions of each IS subtype were calculated using cases with evident and probable determined aetiology in each area (U: urban area; R: rural area). The size of each pie represents area-specific 9-year incidence of total IS, including cases with both determined and undetermined aetiology.

Table 8.2: Baseline characteristics by ischaemic stroke (IS) subtypes.

	CE (n=160)	LAA (n=2,395)	SAO (n=4,888)	Undetermined (Incomplete Investigation) (n=5,276)	No Stroke (n=445,246)
<i>Demographic factors</i>					
Age, mean (s.d.) (years)	61.9 (9.4)	59.9 (8.9)	58.7 (9.5)	59.0 (9.8)	50.9 (10.3)
Female (%)	46.9	46.3	59.1	54.9	59.7
Urban (%)	59.4	51.4	64.4	41.4	43.0
<i>Medical history and health status (%)</i>					
Diabetes	10.6	14.5	12.5	13.3	4.9
Coronary heart disease	14.4	7.8	7.4	7.2	2.2
Hypertension	59.4	61.9	54.6	59.0	30.6
Blood pressure drugs	29.4	24.0	21.5	23.1	9.2
Cardioaortic source of cerebral embolism*	100.0	4.1	0.3	3.4	—
<i>Physical measurements, mean (s.d.)</i>					
SBP (mmHg)	141.9 (25.6)	144.7 (23.8)	139.9 (23.1)	142.9 (24.5)	129.3 (20.2)
DBP (mmHg)	80.8 (13.5)	82.2 (12.6)	81.1 (12.2)	81.2 (12.3)	77.2 (10.8)
BMI (kg m ⁻²)	24.8 (3.8)	24.2 (3.3)	24.6 (3.5)	24.4 (3.5)	23.6 (3.3)
<i>Lifestyle factors (%)</i>					
Male ever-regular smokers	71.8	74.9	71.6	75.8	74.5
Female ever-regular smokers	2.7	4.9	5.7	4.5	2.9
Male ever-regular drinkers	38.8	42.1	44.3	41.5	41.9
Female ever-regular drinkers	2.7	2.2	2.9	2.5	2.9

Evident and probable cases of cardioaortic embolism (CE), large artery atherosclerosis (LAA), and small artery occlusion (SAO) are included.

Hypertension defined as SBP $\geq$ 140 mmHg, DBP $\geq$ 90 mmHg, or use of antihypertensive medication.

Ever-regular smokers include current- and ex-regular smokers.

Ever-regular drinkers include current-regular, ex-regular and reduced-intake alcohol drinkers.

*Cardioaortic sources of cerebral embolism include those with high- and low- primary risk recognised by CCS and were identified from retrieved hospital stroke records rather than at baseline.

Of the IS cases with evident and probable aetiology, 2% had CE, 32% had LAA, and 66% had SAO (Table 8.2). Overall, individuals with IS were older and more likely to be men than women, or have diabetes and hypertension, or a prior history of coronary heart disease (CHD), or have higher levels of body mass index (BMI)

compared those without stroke (Table 8.2). Baseline risk factor measurements were broadly similar for different IS subtypes. However, history of CHD and cardioaortic sources of embolism (e.g., atrial fibrillation [AF]) were more common in individuals with CE, while individuals with SAO were more likely to be women and to live in urban areas. There were also substantial differences in distribution of IS types by geographic regions within China (Figure 8.3), with overall 9-year incidence of IS varying over 6-fold across areas (19.7% in Harbin, 3.2% in Suzhou) and the ratio of SAO-to-LAA cases varying substantially. Notably, CE consistently accounted for only a small proportion of IS in all regions. In contrast, CE had the highest 5-year all-cause mortality and recurrent stroke rates (40.7% and 50.5%, respectively), followed by LAA (17.4% and 45.5%) and SAO (11.1% and 39.4%) (Figure 8.4). Overall, the 5-year all-cause mortality and recurrent stroke rates for IS were 16.7% and 40.7% respectively. Recurrent strokes following each of the IS subtypes were primarily ischaemic (91-94%) with 5-6% being haemorrhagic (Table 8.3).

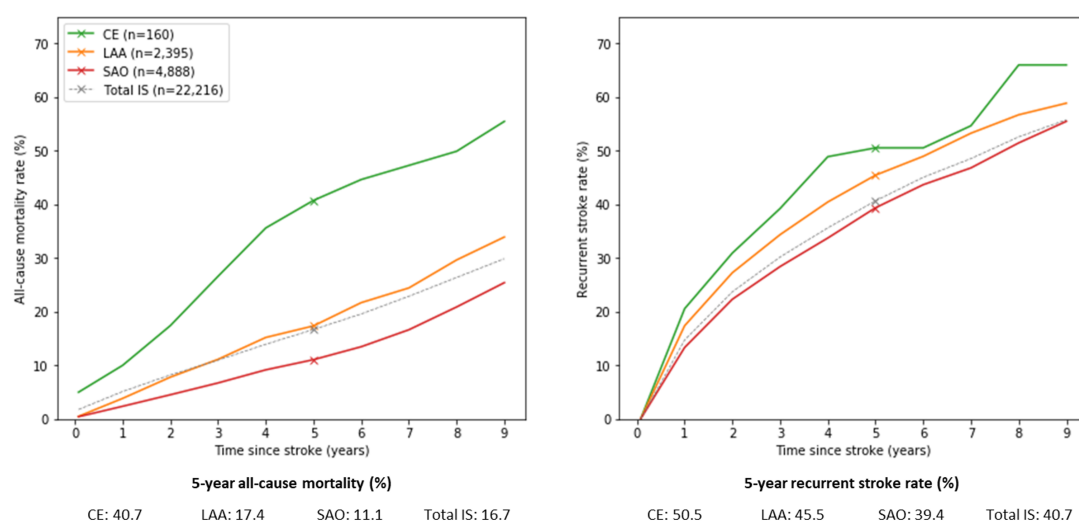


Figure 8.4: Estimated cumulative all-cause mortality and recurrent stroke rates after first ischaemic stroke (IS) of different subtypes. CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion. Evident and probable cases of CE, LAA, and SAO were included. Total IS: total IS including cases with both determined and undetermined aetiology.

Table 8.3: Breakdown of second strokes by pathological types following a first stroke of different ischaemic stroke (IS) subtypes.

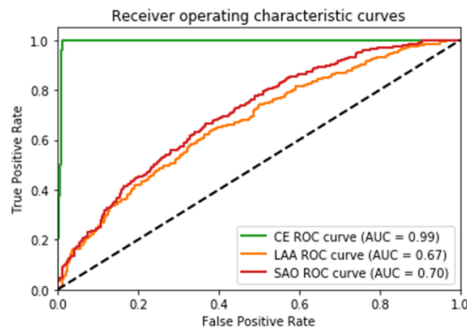
Second Stroke Type	First Ischaemic Stroke Subtype		
	CE	LAA	SAO
	Number of Events (%)	Number of Events (%)	Number of Events (%)
IS (I63)	60 (94%)	766 (91%)	1,506 (91%)
ICH (I61)	4 (6%)	51 (6%)	85 (5%)
SAH (I60)	0 (0%)	0 (0%)	5 (0%)
Unspecified (I64)	0 (0%)	26 (3%)	68 (4%)

CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion. IS: Ischaemic stroke. ICH: Intracerebral haemorrhage. SAH: Subarachnoid haemorrhage.

8.3.2 Classification of Ischaemic Stroke Cases

The multiclass LR model yielded good overall discrimination on the IS cases with evident and probable aetiology in the test set (weighted one-vs-one AUC: 0.80) with near-perfect discrimination of CE (AUC: 0.99), and modest discrimination of LAA (AUC: 0.67) and SAO (AUC: 0.70) (Figure 8.5). The overall accuracy of predictions was 69%. While the model was highly sensitive to SAO (the most common IS subtype), this came at the cost of tending to misclassify LAA as SAO (Figure 8.5). Model performance for the unseen IS cases with possible aetiology was comparable, with slightly lower AUC scores (Figure 8.6). The most important variable for identifying CE was any finding of cardioaortic sources of embolism, while variables important for distinguishing LAA and SAO included region, sex, health coverage, and family history of stroke (Table 8.4).

When applied to incompletely investigated IS cases with undetermined aetiology, the ML-classified strokes (3% CE, 12% LAA, 85% SAO) had broadly similar outcomes to the “true” evident and probable cases of each subtype. ML-classified CE cases had the highest 5-year all-cause mortality and recurrent stroke rates (45.7% and 49.7%, respectively), followed by LAA (33.0% and 43.8%) and SAO (18.3% and 38.8%) (Figure 8.7). However, mortality rates subsequent to a ML-classified incompletely investigated stroke were elevated compared to “true” cases of each IS subtype, and may be partially explained by disparities in health resources

**One-vs-One ROC AUC scores:**

0.81 (macro-averaged),
0.80 (weighted by prevalence)

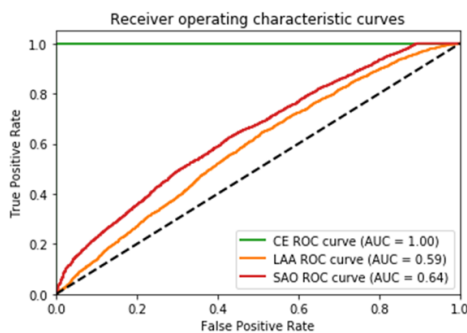
One-vs-Rest ROC AUC scores:

0.79 (macro-averaged),
0.70 (weighted by prevalence)

	ML-Classified CE	ML-Classified LAA	ML-Classified SAO
Actual CE	16	8	0
Actual LAA	8	86	265
Actual SAO	1	67	666

	PPV	Sensitivity	F1 Score	Support
CE	0.64	0.67	0.65	24
LAA	0.53	0.24	0.33	359
SAO	0.72	0.91	0.80	734
Accuracy			0.69	1,117
Macro-Averaged	0.63	0.60	0.59	1,117
Weighted Average	0.66	0.69	0.65	1,117

Figure 8.5: Performance of multiclass logistic regression (LR) model on evident and probable cases of each ischaemic stroke (IS) subtype in test set.

**One-vs-One ROC AUC scores:**

0.76 (macro-averaged),
0.76 (weighted by prevalence)

One-vs-Rest ROC AUC scores:

0.74 (macro-averaged),
0.64 (weighted by prevalence)

	ML-Classified CE	ML-Classified LAA	ML-Classified SAO
Actual CE	150	85	0
Actual LAA	0	274	1679
Actual SAO	0	573	5327

	PPV	Sensitivity	F1 Score	Support
CE	1.00	0.64	0.78	235
LAA	0.29	0.14	0.19	1,953
SAO	0.76	0.90	0.83	5,900
Accuracy			0.71	8,088
Macro-Averaged	0.68	0.56	0.60	8,088
Weighted Average	0.65	0.71	0.67	8,088

Figure 8.6: Performance of multiclass logistic regression (LR) model on possible cases of each ischaemic stroke (IS) subtype in test set.

Table 8.4: Top risk factors for classifying ischaemic stroke (IS) subtypes.

CE		LAA		SAO	
Risk Factor	OR [95% CI]	Risk Factor	OR [95% CI]	Risk Factor	OR [95% CI]
ce_risk	84,607 [2.1*10 ⁴ -2.2*10 ⁶]	Hunan	2.81 [2.21-3.94]	ce_risk	0.13 [0.09-0.17]
Age (10 years)	0.20 [0.05-0.84]	Harbin	0.55 [0.43-0.71]	Hunan	0.34 [0.25-0.44]
BMI	3.12 [1.98-6.11]	Female	0.59 [0.46-0.73]	Harbin	1.92 [1.44-2.45]
Fat percent	0.77 [0.68-0.91]	Suzhou	1.64 [1.17-2.30]	Suzhou	0.57 [0.41-0.85]
		Henan	0.68 [0.47-0.97]	Female	1.74 [1.41-2.31]
		Haikou	1.45 [1.08-1.98]	Henan	1.48 [1.03-2.04]
		Liuzhou	0.74 [0.54-0.99]	Haikou	0.69 [0.50-0.92]
		Health cover	0.79 [0.63-0.97]	Liuzhou	1.35 [1.00-1.82]
		Father stroke	0.86 [0.69-0.97]	Health cover	1.29 [1.08-1.60]
		SBP_noHRX	1.10 [1.03-1.26]	Father stroke	1.17 [1.03-1.45]

CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion. OR: Odds ratio. Risk factors ranked by magnitude of logistic regression coefficient and included if 95% CI indicated statistical significance. Red cells indicate increased likelihood of corresponding IS subtype. Green cells indicate decreased likelihood of corresponding IS subtype.

and quality of care across China. For example, in rural areas, 59% of cases were incompletely investigated while only 48% were incompletely investigated in urban areas. However, these rural areas also had higher 5-year mortality rates following a first IS event (20.0% in rural areas vs 14.2% in urban areas).

Finally, sensitivity analyses removing silent cerebral infarcts generally yielded similar results to the main analyses (Appendix: Tables F.2-F.3, Figures F.1-F.4). However, since CCS classified most silent cerebral infarcts as SAO (Appendix: Table F.2), exclusion of these events lowered the ratio of SAO-to-LAA cases and led the LR model to classify a larger proportion of strokes as LAA (Appendix: Figures F.1-F.4). Silent cerebral infarcts were also observed to have lower 5-year all-cause mortality (10.1%) and recurrent stroke rates (35.1%) than non-silent strokes (Appendix: Figure F.3).

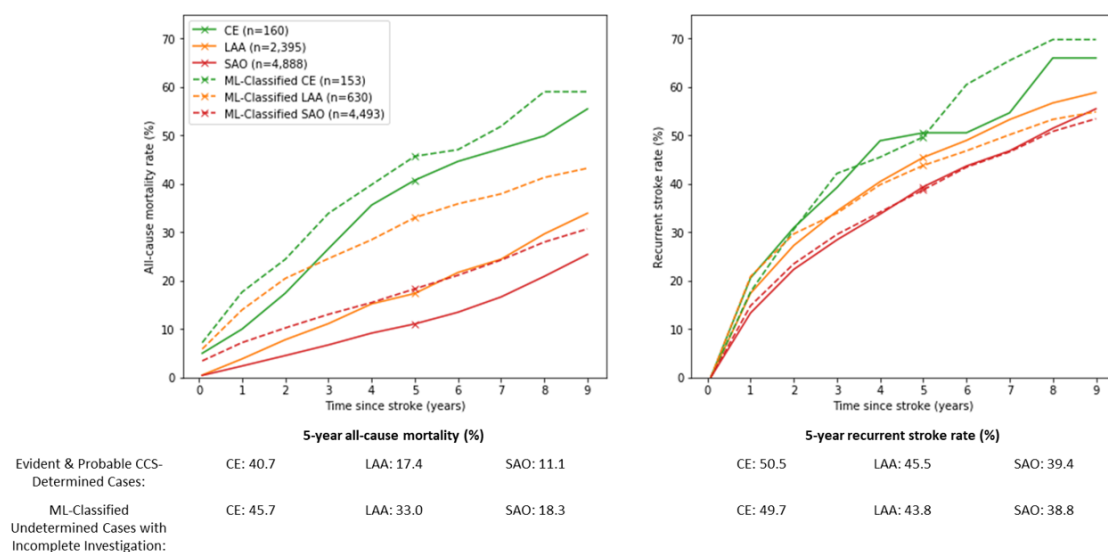


Figure 8.7: All-cause mortality and recurrent stroke rates of different ischaemic stroke (IS) subtypes for (i) cases with evident and probable determined aetiology and (ii) machine learning (ML)-classified strokes with undetermined aetiology due to incomplete investigation.

8.4 Discussion

In this large contemporary study of Chinese adults, over 7% of previously stroke-free participants had a reported primary IS event during nine years of follow up. The prognosis of IS cases was poor, with over 17% of individuals dying within 5 years of a first IS event and over 40% having a recurrent stroke (>90% IS) within 5 years. Moreover, the prognosis differed markedly between subtypes, with CE having the worst outcomes, followed by LAA and SAO. Similar heterogeneity in outcomes had been previously reported in smaller studies in both China and Europe [149, 153].

Compared to a large international consortium of IS studies from the US and Europe with 16,954 IS cases [154], proportions of CE among cases with evident and probable etiology were 18-fold lower (2% vs 36%), proportion of LAA comparable (32% vs 33%), and proportion of SAO 2-fold higher (66% vs 31%) in this Chinese population. These results support findings of lower proportion of CE and higher proportion of SAO in other Chinese populations, and highlight differences in distribution of IS subtypes between individuals of Chinese and European ancestry [27, 148–150]. The low rates of CE in Chinese individuals is consistent with the low

prevalence of CHD in this population and may also be related to underreporting of AF in China, particularly in areas with fewer health resources.

Mean levels of risk factors at baseline were broadly similar across IS subtypes. However, consistent with a systematic review of risk factors for IS subtypes in Chinese populations [148], diabetes was more common in individuals with LAA, CHD and cardioaortic sources of embolism (e.g., AF) were more common in individuals with CE, and diabetes was more common in SAO than in CE (but lower than in those with LAA). In contrast with the systematic review, we did not observe any differences in the prevalence of hypertension between different subtypes of IS.

Finally, we developed a ML-based classifier for probabilistically classifying IS subtypes based on pre-stroke risk factor measurements and demonstrated the model's ability to meaningfully classify IS cases of undetermined aetiology in settings of incomplete investigation. Previous studies of IS subtyping using ML have focused on classifying IS cases from post-stroke clinical findings obtained from electronic health records to match the implementation of gold-standard subtyping algorithms [155–157]. However, by relying on pre-stroke risk factors, the present study was able to classify cases with incomplete investigation, where algorithms such as CCS were unable to assign a single likely aetiology. These undetermined cases, comprising 24% of total primary IS cases in the present study (and especially common in rural areas), enable a major opportunity to further inform stroke care in low-resource settings and could increase the statistical power for some research questions including studies of genetic determinants of IS stroke subtypes. Further work is required to validate the utility of these labels in such settings.

This study had several strengths, including the prospective study design, the large number of participants, detailed data on stroke types, and complete long-term follow-up for both fatal and non-fatal incident stroke cases. However, while CKB was designed to study disease incidence in diverse regions of China, it is not representative of the overall Chinese population and may not be directly generalisable. Second, baseline data on certain established risk factors for stroke such as AF and blood cholesterol levels were not available and, if available, could

have enhanced model performance. Third, the inclusion of region in our classification model may limit the usefulness of this approach in clinical practice. However, since our motivation was to classify IS cases in CKB, where regional data is available and highly informative, we chose to include these variables to maximise model performance. Differences between regions such as ambient temperature [100], are likely to underly the differences in proportions of IS subtypes and should be studied further. Finally, since certain variables for implementing CCS were unavailable, the CCS algorithm had to be slightly adapted prior to fitting in CKB (Appendix F.1).

8.5 Conclusion

The study presented in this chapter provides contemporary evidence for differences in IS subtypes, their risk factor associations, and their outcomes in Chinese adults living in diverse areas within China. Furthermore, based on these differences, a novel approach is proposed for classifying IS cases of undetermined aetiology, potentially informing stroke care and substantially increasing the statistical power of future stroke studies.

9

Conclusions and Future Work

Stroke is a leading cause of death and disability worldwide, and about three-quarters of all stroke cases currently occur in low- and middle-income countries (LMICs), including China [5, 6]. China, in particular, has the largest number of stroke cases in the world and accounts for one-third of global stroke deaths [2]. About half of all strokes could be prevented by adopting population-based lifestyle strategies or wider use of blood pressure-lowering or lipid-lowering medication in high-risk individuals [4]. However, there remain significant obstacles.

First, incidence of stroke and prevalence of its major risk factors vary substantially around the world. Risk prediction models for cardiovascular disease (CVD) and stroke are important tools for identifying high-risk individuals who may benefit from preventative interventions [30, 31]. However, existing models have primarily been developed and validated in cohorts of European ancestry from high income countries (HICs) and may have limited clinical utility in the LMICs that currently bear the brunt of the global stroke burden [8, 46, 47]. Large studies collecting high-quality data on stroke and its risk factors are unavailable for many LMICs, making study of these regions particularly challenging.

Second, stroke is a heterogeneous disease, comprising various pathologies with varying risk factors and prognoses. Ischaemic strokes (IS) differ in important ways from haemorrhagic strokes (HS), and even among IS cases, various aetiological

subtypes exist including cardioaortic embolism (CE), large artery atherosclerosis (LAA), and small artery occlusion (SAO). Existing risk prediction models often treat total CVD or total stroke as a single composite endpoint and do not provide granularity on these pathological types and subtypes of stroke. Such granularity is especially important in LMICs such as China where the proportion of HS is substantially higher than in HICs [6]. However, as one subdivides stroke into its various component pathologies, the challenges of collecting data on a sufficient number of cases becomes further exacerbated.

Given these challenges, the China Kadoorie Biobank (CKB) [78, 79] offers an unprecedented opportunity to study stroke in China. One of the largest and most contemporary prospective cohort studies of its kind, the CKB is a valuable resource providing high-quality and granular stroke data on 0.5 million contemporary Chinese adults, with very few missing values and remarkably low loss to follow-up. While not designed to be representative of the overall Chinese population, the CKB also provides insight into the geographic variance within China, recruiting individuals from 10 diverse areas (5 urban, 5 rural) across the country. Using this resource, the work presented in this thesis makes several contributions that advance our understanding of stroke in China.

9.1 Summary of Thesis

In this thesis we have combined the use of traditional epidemiological approaches and techniques from the field of machine learning (ML) to model and better understand stroke risk in Chinese adults. We began in Chapter 1 with an introduction to stroke in China and laid out the main objectives of the thesis. In Chapter 2, we provided an overview of stroke pathology and primary prevention strategies, with a special focus on stroke risk models. Most risk prediction tools currently used in clinical practice are based on traditional statistical approaches such as Cox regression, and a number of studies have developed such models in various populations. However, only a handful of studies have explored the use of ML approaches for risk prediction, and these models have not been widely adopted in

clinical practice. Our review highlighted important opportunities for advancing the field by studying more contemporary cohorts in LMICs such as China, by developing separate risk prediction models for IS and HS, and by investigating the differences between IS subtypes.

In Chapter 3, we described the CKB, providing details about the overall study design and additional efforts taken to validate and adjudicate the stroke endpoints recorded during follow up. We discussed how we selected the subset of CKB data used throughout this thesis and then conducted descriptive analyses detailing the distributions of well-established stroke risk factors in CKB. We ended the chapter with a discussion of missing data and censoring, highlighting the completeness and limited loss to follow-up in this dataset. While the CKB has some limitations such as representativeness and lack of data on certain established stroke risk factors (e.g., AF, left ventricular hypertrophy, blood lipid levels, etc.), it is an invaluable resource for the study of stroke (and other chronic diseases) in China.

In Chapter 4, we investigated the use of traditional Cox regression approaches for prediction of stroke risk in CKB. We demonstrated the poor calibration of the 2017 Framingham Stroke Risk Profile (FSRP) [8], which substantially underestimated stroke risk in Chinese adults. Then, through recalibration, refitting, area stratification and inclusion of additional risk factors, we developed Cox models that predicted 9-year risks of total stroke with a 6%-8% increase in area under the receiver operating characteristic curve (AUC) and >90% reduction in the calibration χ^2 statistic compared to the FSRP. We also developed separate risk prediction models for IS and HS, which yielded similar performance. The findings from this chapter highlighted the issues of using risk scores from HICs in China, demonstrated the geographic diversity between regions within China, and generated novel risk prediction tools for stroke and its major pathological types that outperformed the FSRP in Chinese adults.

In Chapter 5, we investigated the use of ML approaches for risk prediction of stroke in CKB and compared these to the traditional Cox regression approach. Among Cox regression, random survival forests (RSF), logistic regression (LR),

support vector machines (SVM), gradient boosted trees (GBT), and multilayer perceptrons (MLP), GBT models yielded the best discrimination and calibration for 9-year risk prediction of stroke, providing incremental improvements over the Cox approach. Improvements over the Cox models were particularly evident for binary classification models that predicted stroke at narrower intervals of follow-up (0-3 years, 3-6 years, and 6-9 years from baseline). However, understanding the strong preference of clinicians for the Cox model-based approaches currently used in clinical practice, we developed an ensemble approach that relied, by default, on the outputs of a traditional Cox model for risk classification and only overrode the classification of the Cox model with that of the GBT model in 1%-2% of individuals. This ensemble approach yielded higher accuracy, specificity, and precision than any of the single model approaches. Moreover, it was intentionally designed as an incremental change to current practice, which could help to facilitate more widespread use and trust of ML methods for risk prediction in clinical practice.

While current risk prediction models generate risk predictions based on measurements of risk factors recorded at a single baseline visit, in Chapter 6, we investigated if risk prediction could be improved from using repeated measurements of time-dependent risk factors including blood pressure (BP), body weight, and lifestyle factors. Using longitudinal resurvey data from CKB individuals with risk factor measurements recorded at three separate visits, we (i) examined changes in CVD risk factors and 3-year absolute risks of stroke in Chinese adults from 2004-2014 and (ii) compared the utility of sequential versus single measurements of CVD risk factors for prediction of stroke in a contemporary study of Chinese adults. We found that individuals at highest risk for stroke in China from 2013-2014 were largely the same individuals who had the fastest growing risk of stroke throughout the previous decade (a phenomenon in epidemiology referred to as the “horse-racing effect”). Due to this effect, measurements of CVD risk factors for prediction of stroke at three separate surveys had limited additional utility compared to those recorded at a single survey. These results provide further support for current guidelines for clinical practice, which suggest that stroke prediction is best estimated and

cost-effectively implemented using models involving measurements of major risk factors recorded at a single visit.

In Chapter 7, we turned our attention toward the application of stroke risk models for informing treatment decisions and addressed an important issue that has been a major obstacle to the wider use of lipid-lowering statins in China. Statin use has increased worldwide from primary prevention of CVD [139], but its use remains low in China [138], partially because of concerns about increasing risk of HS. While the benefits of statins for reducing risk of IS and major coronary events (MCEs) certainly outweigh the potential hazards of increasing risk of HS on a population level, it was unknown if there were particular individuals with high risk of HS for whom statin therapy could be contraindicated. We developed two approaches (i.e., a Cox-based approach and ML-based GBT approach) for effectively identifying the individuals whose first stroke event is most likely to be HS, and demonstrated that the benefits of statins outweigh the risks in nearly all (>97%) Chinese adults. We estimated that by lowering LDL-C by 1 mmol l^{-1} with statins in CKB, about 150 events (IS, HS, and MCEs) and 12 deaths could have been avoided per 10,000 Chinese adults over the nine-year follow-up period, with even greater effect in older individuals with higher vascular risk. These findings suggest that the fear of unjustifiably increasing risk of HS in Chinese adults is largely unfounded and provide support for wider use of statins, as advocated by both recent Chinese and international guidelines [102, 118, 119, 137].

In Chapter 8, we studied differences between aetiological IS subtypes in CKB and developed a ML approach for probabilistically classifying IS cases of undetermined aetiology due to incomplete investigations. We characterised the heterogeneous distributions, risk factors, and outcomes between IS subtypes, and demonstrated important differences between IS subtypes in Chinese and European populations. Furthermore, while gold-standard algorithms for classification of IS subtypes such as the Causative Classification System for Ischemic Stroke (CCS) frequently fail to identify a single likely aetiology in settings with incomplete investigations (e.g., missing brain imaging, cardiac tests, vascular tests, etc.), we developed a multiclass

LR classifier for resolving incompletely investigated cases left unclassified by CCS (24% of IS cases in CKB) using risk factor data. We validated our classification by showing that the distribution, recurrence, and mortality of IS subtypes among cases with undetermined aetiology that were ascertained by probabilistic ML classification were comparable to those classified after clinical adjudication using CCS. Classifying these IS cases of undetermined aetiology can potentially inform stroke care and substantially increase the statistical power of future stroke studies.

9.1.1 Key Contributions

The key contributions of this thesis therefore consist of:

- A review of traditional statistical approaches and machine learning approaches for stroke risk prediction
- Demonstration of the poor calibration of the FSRP in Chinese adults
- Development of novel risk prediction models for total stroke (using traditional, ML, and ensemble approaches) that outperform the FSRP in Chinese adults
- Development of separate risk prediction models for IS and HS
- Evidence that sequential measurements of CVD risk factors (recorded during 3 visits over a 10-year period) have limited additional predictive utility compared to a recent set of measurements recorded at a single visit
- Application of risk prediction models for IS and HS to demonstrate that concerns about increasing risk of HS should not preclude the wider use of statin therapy in China
- Characterisation of the heterogeneity of IS subtypes (CE, LAA, and SAO), including a comparison between Chinese and European populations
- Development of a ML approach for probabilistically classifying incompletely investigated IS cases of undetermined aetiology into IS subtypes

9.2 Future Work

9.2.1 External Validation and Updating of Models

Prior to implementing any of the models presented in this thesis in clinical practice, external validation in independent populations in China and potentially other LMICs is required. However, in China, there is currently a lack of contemporary regional cohorts of middle-aged and older adults with sufficiently large sample size and follow-up of over 9 years to conduct this validation. Since the mid-2010s, new regional cohorts have been initiated under the China Precision Medicine Initiative [158] and regional electronic resident health records system. As the follow-up data comes in from these cohorts, there will be new opportunities for externally validating the models developed in this thesis. Moreover, as is true with any risk prediction model, it will be important to continually update our models to reflect temporal changes in China, and these new sources of data will play an important role in this process. To allow other researchers to implement and validate the risk prediction models developed in this thesis (particularly those presented in Chapters 4 and 5), we have made the statistical code for our models publicly available at: <https://github.com/ckbiobank/ckb-stroke-risk-models>.

9.2.2 Validation of ML-Classified Ischaemic Stroke Subtype Endpoints

Further validation of the IS subtype endpoints provided by our ML approach in Chapter 8 is also required. For example, using these endpoints, studies can be performed to test if well-known genetic determinants of CE, LAA, and SAO can be replicated among these newly-classified cases of previously undetermined aetiology. Once sufficiently validated, these IS cases (making up 24% of IS cases in CKB) can be included in future analyses and contribute to substantially greater statistical power, for example, to discover new genetic determinants of each IS subtype.

9.2.3 Translation to Clinical Practice

The ultimate goal of all of the work presented in this thesis is to meaningfully impact stroke prevention in China. While our risk prediction models cannot be immediately implemented without external validation, our findings can still play a significant role in influencing the direction of future work in a way that will have lasting effects on clinical practice.

For example, the risk prediction models presented in Chapters 4 and 5 highlight the substantial geographic variance in IS and HS incidence within China. Thus, while other international guidelines have typically advocated the use of a single equation nationally (e.g., the Pooled Cohort Equations [46] in the United States and QRISK3 [47] in the United Kingdom), it may be particularly important to develop regional stroke or CVD scores in China. Current work is being performed by our collaborators to integrate cross-sectional data on risk factors and national mortality statistics collected by the Chinese Center for Disease Control and Prevention to develop a new CVD risk score for the whole of China. However, informed by the findings of this thesis, efforts are now being made to obtain more granular regional mortality statistics to adjust for these important regional differences. This change may lead to a CVD model that is ultimately much more useful for clinical practice in diverse areas across China.

As another example, in Chapter 7, we presented findings that support a wider use of statins in China and estimated the number of events and deaths that could be avoided among individuals of different ages and vascular risk. We hope that these findings will generally alleviate concerns about the effect of statins on HS risk and contribute to meaningful improvements to stroke prevention in the short-term. However, in the longer-term, we believe that future research should incorporate our estimates into economic studies that weigh the benefits and costs of reducing events and deaths in each age and risk group as well as the expected number of quality-adjusted life years (QALYs) gained from preventing each event. We believe that such analysis can lead to better clinical guidelines for stroke prevention

than the current paradigm of setting arbitrary treatment thresholds based on absolute 10-year risks of CVD.

9.3 Closing Thoughts

As the global burden of stroke continues to shifted towards LMICs such as China, continual work will need to focus on these countries. New datasets will need to be collected and new models will need to be developed. Even in HICs, risk prediction of stroke and CVD is not a solved problem, as models will need to be continually updated to adjust for temporal trends.

Within the field of CVD risk prediction, exciting new developments such as polygenic risk scores [159] may improve risk prediction and open up new opportunities for prevention of CVD – the leading cause of death in the world. However, when designing risk prediction tools for LMICs, such benefits need to be weighed against the practical considerations of what risk factor information can be reasonably obtained in lower resource settings. For example, in China, lipid testing is only free for people ≥ 65 years old, and as a consequence, models that rely on cholesterol levels are substantially limited in utility among those aged < 65 years. As researchers continue to discover new risk factors for stroke, risk prediction models for LMICs will still need to be designed in a way that balances predictive performance with widespread applicability.

Finally, while the current paradigm of CVD risk prediction has focused on absolute risk of events and mortality, further consideration needs to be given in the future to how these models are used and communicated in practice. Many of the risk scores currently used in clinical practice are dominated by the effects of age and therefore target interventions toward the oldest individuals who may have the fewest QALYs to gain. These models may also fail to communicate risk effectively to younger individuals who have low absolute risk due to their age, but who are at significantly higher risk than their peers. As data on QALYs becomes increasingly available, we expect existing paradigms of CVD prevention to continue to evolve.

Overall, improvements in the prediction of CVD and stroke presents an exciting opportunity to save numerous lives around the world, especially in low- and middle-income countries. The contributions of this thesis take a small step toward this future, providing important insights into the risk of stroke and its various pathologies in China – an understudied region that currently bears the greatest burden of stroke in the world.

Appendices



CKB Risk Factor Indicators and Adjudication Variables

Table A.1: CKB risk factor indicators.

Name	Definition/Corresponding CKB Survey Question
Risk Factor Indicators Included in FSRP	
Sex	Sex (0:male, 1: female)
Age (10 yrs)	Age in decades
Current Smoking	Do you currently smoke? (0: no, 1: yes)
CHD	Has a doctor EVER told you that you had coronary heart disease? (0: no, 1: yes)
Age 65+	Age ≥ 65 years? (0: no, 1: yes)
DM, if Age <65	IF age < 65 years, do you have diabetes? (0: no, 1: yes)
DM, if Age 65+	IF age < 65 years, do you have diabetes? (0: no, 1: yes)
HTN Rx	Do you use blood pressure drugs? (0: no, 1: yes)
SBP per 10 mmHg, if no HRx	IF not using blood pressure drugs, what is your systolic blood pressure (in mmHg) / 10
SBP per 10 mmHg, if HRx	IF using blood pressure drugs, what is your systolic blood pressure (in mmHg) / 10
Other CKB Risk Factor Indicators	
Gansu	Recruited from Gansu? (0:no, 1: yes)
Haikou	Recruited from Haikou? (0:no, 1: yes)
Harbin	Recruited from Harbin? (0:no, 1: yes)
Henan	Recruited from Henan? (0:no, 1: yes)
Hunan	Recruited from Hunan? (0:no, 1: yes)
Liuzhou	Recruited from Liuzhou? (0:no, 1: yes)
Qingdao	Recruited from Qingdao? (0:no, 1: yes)
Sichuan	Recruited from Sichuan? (0:no, 1: yes)
Suzhou	Recruited from Suzhou? (0:no, 1: yes)
Zhejiang	Recruited from Zhejiang? (0:no, 1: yes)
region_is_urban	Is region urban? (0: no, 1: yes)
sbp_mean	Systolic blood pressure in mmHg (mean of two measurements)
has_diabetes	Participant has history of diabetes (reported OR random blood glucose ≥ 11.1 mmol/L OR fasting blood glucose ≥ 7.0 mmol/L) (0: no, 1: yes)
household_size	How many people live together in your household?

has_health_cover	Do you have health care coverage? (0: no, 1: yes)
years_since_quitting_smoking	How many years since quitting smoking? (0 if you currently smoke or never smoked; decimal values allowed)
has_copd	Has a doctor EVER told you that you had COPD? (0: no, 1: yes)
hypertension_diag	Has a doctor EVER told you that you had hypertension? (0: no, 1: yes)
rheum_heart_dis_diag	Has a doctor EVER told you that you had rheumatic heart disease? (0: no, 1: yes)
tb_diag	Has a doctor EVER told you that you had TB? (0: no, 1: yes)
cirrhosis_hep_diag	Has a doctor EVER told you that you had cirrhosis/chronic hepatitis? (0: no, 1: yes)
peptic_ulcer_diag	Has a doctor EVER told you that you had a peptic ulcer? (0: no, 1: yes)
gall_diag	Has a doctor EVER told you that you had a gallstone/gallbladder dis.? (0: no, 1: yes)
asthma_diag	Has a doctor EVER told you that you had asthma? (0: no, 1: yes)
kidney_dis_diag	Has a doctor EVER told you that you had kidney disease? (0: no, 1: yes)
fracture_diag	Has a doctor EVER told you that you had a fracture? (0: no, 1: yes)
rheum_arthritis_diag	Has a doctor EVER told you that you had rheumatoid arthritis? (0: no, 1: yes)
neurasthenia_diag	Has a doctor EVER told you that you had neurasthenia? (0: no, 1: yes)
head_injury_diag	Has a doctor EVER told you that you had a head injury? (0: no, 1: yes)
cancer_diag	Has a doctor EVER told you that you had cancer? (0: no, 1: yes)
blood_transfusions	How many blood transfusions have you received? (if none, put 0)
children	How many children do you have?
siblings	How many siblings do you have?
mother_still_alive	Is your mother still alive? (0: no, 1: yes)
father_still_alive	Is your father still alive? (0: no, 1: yes)
mother_stroke	Did your mother ever have a stroke? (0: no, 1: yes)
mother_heart_attack	Did your mother ever have a heart attack? (0: no, 1: yes)
mother_diabetes	Did your mother ever have diabetes? (0: no, 1: yes)
mother_cancer	Did your mother ever have cancer? (0: no, 1: yes)
father_stroke	Did your father ever have a stroke? (0: no, 1: yes)
father_heart_attack	Did your father ever have a heart attack? (0: no, 1: yes)
father_diabetes	Did your father ever have diabetes? (0: no, 1: yes)
father_cancer	Did your father ever have cancer? (0: no, 1: yes)
siblings_stroke	Did your siblings (incl. half siblings) ever have a stroke? (0: if no, else record the number with disease)
siblings_heart_attack	Did your siblings (incl. half siblings) ever have a heart attack? (0: if no, else record the number with disease)
siblings_diabetes	Did your siblings (incl. half siblings) ever have diabetes? (0: if no, else record the number with disease)
siblings_cancer	Did your siblings (incl. half siblings) ever have cancer? (0: if no, else record the number with disease)
children_stroke	Did your children ever have a stroke? (0: if no, else record the number with disease)
children_heart_attack	Did your children ever have a heart attack? (0: if no, else record the number with disease)
children_diabetes	Did your children ever have diabetes? (0: if no, else record the number with disease)
children_cancer	Did your children ever have cancer? (0: if no, else record the number with disease)
met	Total daily physical activity (Metabolic Equivalent of Task [MET hours/day])
met_hours	Total daily hours spent on physical activity
standing_height_cm	Standing height in cm (without shoes)
sitting_height_cm	Sitting height in cm
waist_cm	Waist measurement in cm
waist_hip_ratio_percent	Waist to hip ratio
weight_kg	Weight in kg (without shoes, but in light clothing)
bmi_calc	BMI calculated from measured height and weight (in kg/m ²)
fat_percent	Body fat percentage
dbp_mean	Diastolic BP in mmHg (mean of two measurements)
heart_rate_mean_10s	Heart rate (beats per minute/10)

smoking_category_1	Does the following characterise your smoking behaviour? – Never smoker (0: no, 1: yes)
smoking_category_2	Does the following characterise your smoking behaviour? – Occasional smoker (0: no, 1: yes)
smoking_category_3	Does the following characterise your smoking behaviour? – Ex regular smoker (0: no, 1: yes)
smoking_category_4	Does the following characterise your smoking behaviour? – Regular smoker (0: no, 1: yes)
emph_bronc_diag	Has a doctor EVER told you that you had emphysema/bronchitis? (0: no, 1: yes)
psych_disorder_diag	Has a doctor EVER told you that you have a psychiatric disorder? (0: no, 1: yes)
highest_education_0	What is the highest level of school education you ever received? – No formal school (0: no, 1: yes)
highest_education_1	What is the highest level of school education you ever received? – Primary school (0: no, 1: yes)
highest_education_2	What is the highest level of school education you ever received? – Middle school (0: no, 1: yes)
highest_education_3	What is the highest level of school education you ever received? – High school (0: no, 1: yes)
highest_education_4	What is the highest level of school education you ever received? – Technical school / college (0: no, 1: yes)
highest_education_5	What is the highest level of school education you ever received? – University (0: no, 1: yes)
occupation_0	What is your current occupation? – Agriculture and related (0: no, 1: yes)
occupation_1	What is your current occupation? – Factory worker (0: no, 1: yes)
occupation_2	What is your current occupation? – Administrator / manager (0: no, 1: yes)
occupation_3	What is your current occupation? – Professional / technical (0: no, 1: yes)
occupation_4	What is your current occupation? – Sales and service (0: no, 1: yes)
occupation_5	What is your current occupation? – Retired (0: no, 1: yes)
occupation_6	What is your current occupation? – House wife / husband (0: no, 1: yes)
occupation_7	What is your current occupation? – Self-employed (0: no, 1: yes)
occupation_8	What is your current occupation? – Unemployed (0: no, 1: yes)
occupation_9	What is your current occupation? – Other or not stated (0: no, 1: yes)
household_income_0	What is the total income last year in your household? – <2,500 yuan (0: no, 1: yes)
household_income_1	What is the total income last year in your household? – 2,500-4,999 yuan (0: no, 1: yes)
household_income_2	What is the total income last year in your household? – 5,000-9,999 yuan (0: no, 1: yes)
household_income_3	What is the total income last year in your household? – 10,000-19,999 yuan (0: no, 1: yes)
household_income_4	What is the total income last year in your household? – 20,000-34,999 yuan (0: no, 1: yes)
household_income_5	What is the total income last year in your household? – ≥35,000 yuan (0: no, 1: yes)
alcohol_category_1	Does the following characterise your alcohol intake? – Never regular (0: no; 1: yes)
alcohol_category_2	Does the following characterise your alcohol intake? – Ex-regular (0: no; 1: yes)
alcohol_category_3	Does the following characterise your alcohol intake? – Occasional (0: no; 1: yes)
alcohol_category_4	Does the following characterise your alcohol intake? – Monthly (0: no; 1: yes)
alcohol_category_5	Does the following characterise your alcohol intake? – Reduced intake (0: no; 1: yes)
alcohol_category_6	Does the following characterise your alcohol intake? – Weekly (0: no; 1: yes)
smoking_now_0	Does the following characterise your smoking behaviour? – Do not smoke now (0: no; 1: yes)
smoking_now_1	Does the following characterise your smoking behaviour? – Only occasionally (0: no; 1: yes)
smoking_now_2	Does the following characterise your smoking behaviour? – Yes, on most days (0: no; 1: yes)

smoking_now_3	Does the following characterise your smoking behaviour? – Yes, daily or almost every day (0: no; 1: yes)
self_rated_health_0	How would you rate your current general health status? – Excellent (0: no, 1: yes)
self_rated_health_1	How would you rate your current general health status? – Good (0: no, 1: yes)
self_rated_health_2	How would you rate your current general health status? – Fair (0: no, 1: yes)
self_rated_health_3	How would you rate your current general health status? – Poor (0: no, 1: yes)
comparative_health_0	How would rate your current general health status compared to someone of your own age? – Better (0: no, 1: yes)
comparative_health_1	How would rate your current general health status compared to someone of your own age? – About the same (0: no, 1: yes)
comparative_health_2	How would rate your current general health status compared to someone of your own age? – Worse (0: no, 1: yes)
comparative_health_3	How would rate your current general health status compared to someone of your own age? – Don't know (0: no, 1: yes)
diet_freq_rice_0	During the past 12 months, about how often did you eat rice? – Daily (0: no, 1: yes)
diet_freq_rice_1	During the past 12 months, about how often did you eat rice? – 4-6 days per week (0: no, 1: yes)
diet_freq_rice_2	During the past 12 months, about how often did you eat rice? – 1-3 days per week (0: no, 1: yes)
diet_freq_rice_3	During the past 12 months, about how often did you eat rice? – Monthly (0: no, 1: yes)
diet_freq_rice_4	During the past 12 months, about how often did you eat rice? – Never/rarely (0: no, 1: yes)
diet_freq_wheat_0	During the past 12 months, about how often did you eat wheat? – Daily (0: no, 1: yes)
diet_freq_wheat_1	During the past 12 months, about how often did you eat wheat? – 4-6 days per week (0: no, 1: yes)
diet_freq_wheat_2	During the past 12 months, about how often did you eat wheat? – 1-3 days per week (0: no, 1: yes)
diet_freq_wheat_3	During the past 12 months, about how often did you eat wheat? – Monthly (0: no, 1: yes)
diet_freq_wheat_4	During the past 12 months, about how often did you eat wheat? – Never/rarely (0: no, 1: yes)
diet_freq_other_staple_0	During the past 12 months, about how often did you eat staple foods other than rice and wheat (corn, millet etc.)? – Daily (0: no, 1: yes)
diet_freq_other_staple_1	During the past 12 months, about how often did you eat staple foods other than rice and wheat (corn, millet etc.)? – 4-6 days per week (0: no, 1: yes)
diet_freq_other_staple_2	During the past 12 months, about how often did you eat staple foods other than rice and wheat (corn, millet etc.)? – 1-3 days per week (0: no, 1: yes)
diet_freq_other_staple_3	During the past 12 months, about how often did you eat staple foods other than rice and wheat (corn, millet etc.)? – Monthly (0: no, 1: yes)
diet_freq_other_staple_4	During the past 12 months, about how often did you eat staple foods other than rice and wheat (corn, millet etc.)? – Never/rarely (0: no, 1: yes)
bowel_movement_freq_0	About how often do you have bowel movements each week? – More than once on most days (0: no, 1:yes)
bowel_movement_freq_1	About how often do you have bowel movements each week? – About daily (0: no, 1:yes)
bowel_movement_freq_2	About how often do you have bowel movements each week? – Once every 2-3 days (0: no, 1:yes)
bowel_movement_freq_3	About how often do you have bowel movements each week? – Less than 3 times a week (0: no, 1:yes)
gum_bleed_freq_0	How often do your gums bleed when you brush your teeth? – Occasionally, rarely or never (0: no, 1:yes)
gum_bleed_freq_1	How often do your gums bleed when you brush your teeth? – Sometimes (0: no, 1:yes)
gum_bleed_freq_2	How often do your gums bleed when you brush your teeth? – Always (0: no, 1:yes)
gum_bleed_freq_3	How often do your gums bleed when you brush your teeth? – Brush teeth rarely or never (0: no, 1:yes)

Missing Value Risk Factor Indicators

missing_mother_history	Missing reported history of stroke, heart attack, diabetes, and/or cancer in mother
missing_father_history	Missing reported history of stroke, heart attack, diabetes, and/or cancer in father
missing_siblings_history	Missing reported history of stroke, heart attack, diabetes, and/or cancer in siblings

Table A.2: CKB stroke adjudication variables.

Name	Definition
imaging_findings	Brain imaging findings (“Lacunar”, “Ischaemic”, “Haemorrhagic”, “Indeterminate”, “ICH Transformation After IS”, “Mainly HS lesion with IS changes”, “Old Lesion”, “Subarachnoid”, “Normal Imaging”, “No Imaging”, “Other”)
intra_extra_test_done	Intracranial / extracranial vascular test done (“Neither”, “Intra Only”, “Extra Only”, “Both Intra and Extra”)
intra_extra_is_abnormal	Any stenosis in intracranial / extracranial artery (“Neither”, “Intra Only”, “Extra Only”, “Both Intra and Extra”, “NULL - No test done”)
intra_extra_loc_art_dgre_stsis	Any stenosis $\geq 50\%$ in intracranial / extracranial artery (“Neither”, “Intra Only”, “Extra Only”, “Both Intra and Extra”, “NULL - No test done”)
ultrasound_plaq_featre	IF plaque, stable (hyperechoic) or unstable (hypoechoic or mixed echoic) - extracranial artery only (“Stable”, “Unstable”, “NULL”)
echo_ecg_test_done	Echocardiogram / ECG done (“Neither”, “ECG Only”, “Echo Only”, “Both ECG and Echo”)
ce_risk	Presence of high-risk / low- or uncertain-risk embolic source as defined by CCS (2:“High-Risk”, 1:“Low-Risk”, 0:“None”, NULL)
systemic_embolization	Systemic embolization (1:Yes, 0:No, NULL)
lacunar_syndrome	Presence of specific clinical lacunar syndromes, <i>e.g.</i> , pure motor, pure sensory, ataxic hemiparesis, clumsy hand (1:Yes, 0:No)

B

Chapter 4 Supplementary Materials

B.1 Testing of Proportional Hazards Assumption

The Proportional Hazards (PH) assumption for the FSRP inputs was checked using the `CoxPHFitter.check_assumptions` method implemented by the `lifelines` package version 0.21.1 [95] in Python version 3.7.0. This method performs a statistical test to test for any time-varying coefficients, and provides visual plots of the scaled Schoenfeld residuals presented against four time transformations for any risk factor that violates the PH assumption. In each plot, a fitted lowess is also presented, along with 10 bootstrapped lowess lines. Deviations of the lowess line from a constant value are violations of the PH assumption.

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

treatment ($p < 0.005$), and systolic blood pressure for individuals with hypertension treatment ($p = 0.04$).

It is important to note that with a large sample size, such as in CKB, even very small violations of the PH assumption would test as statistically significant. To observe the impact of this effect, we repeated tests of the PH assumption for a randomly selected 10% of CKB men and women (17,449 men; 25,376 women). In men, only systolic blood pressure for individuals without hypertension treatment ($p = 0.02$) and systolic blood pressure for individuals with hypertension treatment ($p = 0.01$) were still identified to violate the PH assumption. However, even among these risk factors, the Schoenfeld residual plots (Figure B.1) showed a lowess line with very minor deviation from a constant value.

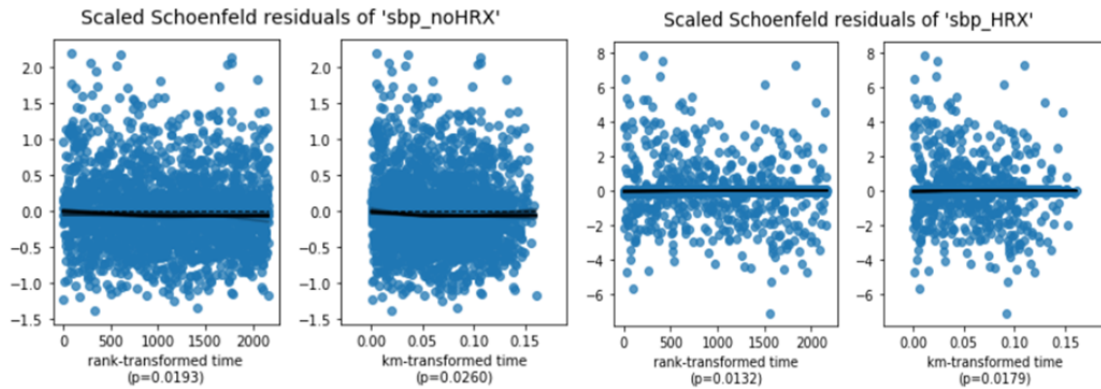


Figure B.1: Schoenfeld residual plots for systolic blood pressure for individuals without hypertension treatment (left 2 subplots) and systolic blood pressure for individuals with hypertension treatment (right 2 subplots) among CKB men.

In women, only anti-hypertensive treatment ($p = 0.02$) and systolic blood pressure for individuals without hypertension treatment ($p = 0.03$) were still identified to violate the PH assumption. Once again, the Schoenfeld residual plots for these risk factors (Figure B.2) showed a lowess line with very minor deviation from a constant value. The results of these tests suggest that while the hazards for every FSRP risk factor may not be perfectly proportional, the PH assumption may still be appropriate in this setting.

Finally, we note that while the PH assumption is important for causal inference, satisfaction of the PH assumption is not necessary for risk prediction applications,

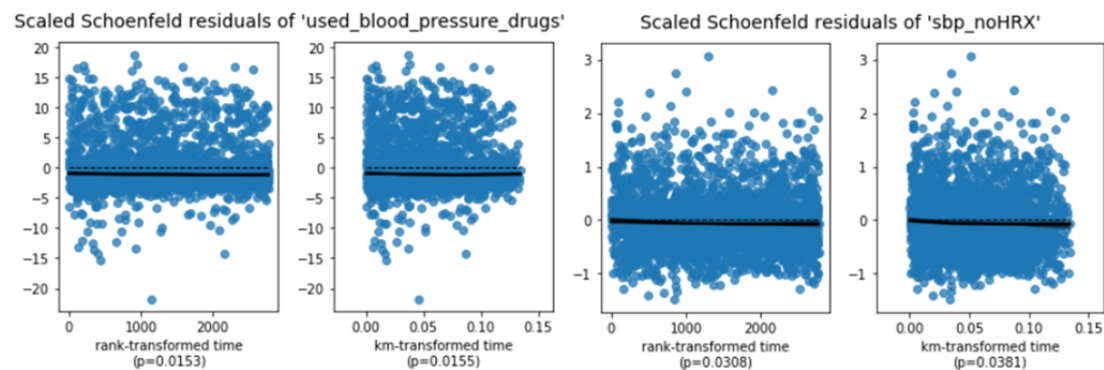


Figure B.2: Schoenfeld residual plots for anti-hypertensive treatment (left 2 subplots) and systolic blood pressure for individuals without hypertension treatment (right 2 subplots) among CKB women.

where the objectives are optimal risk discrimination and risk calibration irrespective of how individual predictions are generated.

B.2 Additional Figures and Tables

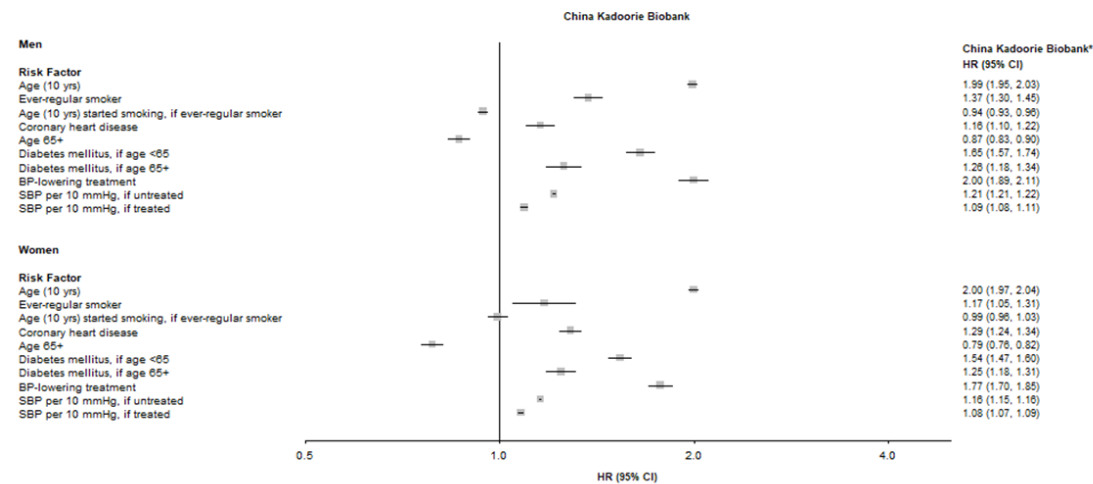


Figure B.3: Multivariable hazard ratios (HRs) and 95% CI for total stroke using Framingham (“2017 FSRP” models) versus CKB (“+Area Stratification” models, modified to include ever-regular smoking and age started smoking) coefficients in men and women.

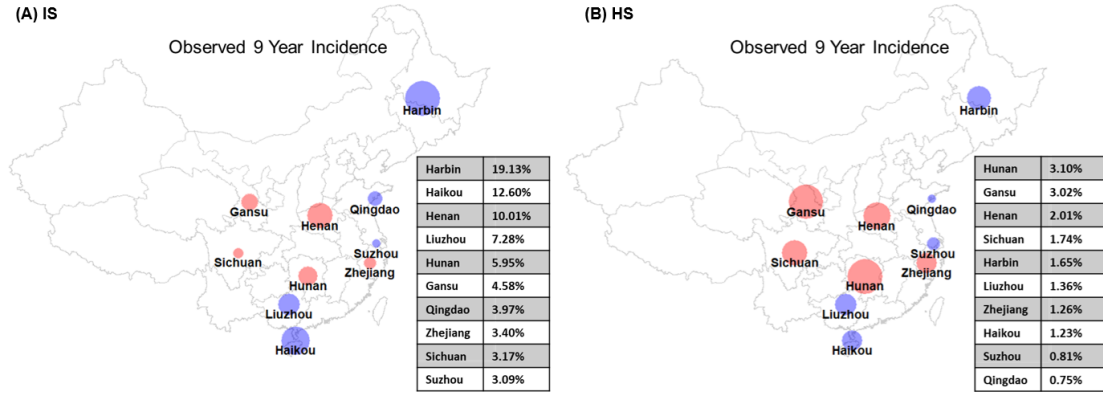


Figure B.4: 9 year incidence for (A) IS and (B) HS by CKB area (adjusted for competing risk of death).

Table B.1: Area-specific discrimination performance for total stroke and stroke pathological types in men and women using “+Area Stratification” models with FSRP inputs.

	Men			Women		
	Total Stroke AUC [95%CI]	IS AUC [95%CI]	HS AUC [95%CI]	Total Stroke AUC [95%CI]	IS AUC [95%CI]	HS AUC [95%CI]
Qingdao	0.84 [0.79-0.89]	0.85 [0.81-0.89]	0.81 [0.71-0.90]	0.81 [0.78-0.84]	0.79 [0.75-0.83]	0.78 [0.71-0.85]
Harbin	0.75 [0.73-0.77]	0.75 [0.73-0.77]	0.77 [0.73-0.81]	0.76 [0.74-0.77]	0.76 [0.74-0.77]	0.80 [0.76-0.84]
Haikou	0.80 [0.77-0.83]	0.80 [0.77-0.83]	0.79 [0.70-0.87]	0.80 [0.77-0.82]	0.77 [0.74-0.79]	0.85 [0.78-0.90]
Suzhou	0.82 [0.79-0.85]	0.82 [0.80-0.85]	0.72 [0.62-0.80]	0.84 [0.81-0.87]	0.82 [0.78-0.85]	0.74 [0.62-0.84]
Liuzhou	0.81 [0.79-0.83]	0.77 [0.75-0.80]	0.84 [0.79-0.89]	0.82 [0.79-0.84]	0.79 [0.77-0.82]	0.75 [0.69-0.81]
Sichuan	0.82 [0.79-0.85]	0.82 [0.78-0.86]	0.86 [0.83-0.89]	0.82 [0.79-0.84]	0.81 [0.78-0.84]	0.86 [0.82-0.90]
Gansu	0.80 [0.77-0.83]	0.75 [0.70-0.79]	0.86 [0.82-0.89]	0.79 [0.77-0.82]	0.77 [0.74-0.80]	0.86 [0.84-0.89]
Henan	0.78 [0.75-0.80]	0.77 [0.75-0.79]	0.74 [0.69-0.78]	0.77 [0.75-0.79]	0.77 [0.75-0.79]	0.77 [0.73-0.81]
Zhejiang	0.81 [0.78-0.84]	0.79 [0.75-0.83]	0.79 [0.74-0.83]	0.82 [0.79-0.85]	0.82 [0.80-0.85]	0.79 [0.73-0.85]
Hunan	0.80 [0.78-0.82]	0.78 [0.75-0.80]	0.81 [0.77-0.84]	0.76 [0.73-0.78]	0.76 [0.73-0.79]	0.79 [0.74-0.83]

Table B.2: Area-specific discrimination performance for total stroke and stroke pathological types in men and women using “+Additional Risk Factors” models.

	Men			Women		
	Total Stroke (66 Selected Risk Factors) AUC [95%CI]	IS (62 Selected Risk Factors) AUC [95%CI]	HS (42 Selected Risk Factors) AUC [95%CI]	Total Stroke (70 Selected Risk Factors) AUC [95%CI]	IS (80 Selected Risk Factors) AUC [95%CI]	HS (38 Selected Risk Factors) AUC [95%CI]
Qingdao	0.84 [0.79-0.88]	0.86 [0.81-0.89]	0.84 [0.78-0.90]	0.82 [0.79-0.85]	0.80 [0.76-0.83]	0.80 [0.73-0.86]
Harbin	0.75 [0.73-0.77]	0.77 [0.75-0.79]	0.78 [0.73-0.83]	0.76 [0.75-0.78]	0.76 [0.75-0.78]	0.81 [0.76-0.86]
Haikou	0.79 [0.75-0.82]	0.81 [0.78-0.84]	0.79 [0.70-0.86]	0.80 [0.78-0.82]	0.77 [0.75-0.79]	0.84 [0.77-0.90]
Suzhou	0.83 [0.80-0.86]	0.83 [0.80-0.86]	0.77 [0.67-0.86]	0.85 [0.82-0.87]	0.82 [0.78-0.86]	0.75 [0.63-0.85]
Liuzhou	0.83 [0.80-0.85]	0.78 [0.76-0.81]	0.84 [0.80-0.88]	0.82 [0.80-0.84]	0.80 [0.77-0.82]	0.74 [0.67-0.81]
Sichuan	0.82 [0.79-0.85]	0.82 [0.78-0.86]	0.88 [0.84-0.91]	0.82 [0.79-0.85]	0.82 [0.79-0.85]	0.86 [0.81-0.90]
Gansu	0.81 [0.78-0.83]	0.77 [0.74-0.81]	0.85 [0.82-0.89]	0.80 [0.77-0.83]	0.78 [0.75-0.81]	0.87 [0.84-0.90]
Henan	0.79 [0.77-0.82]	0.77 [0.75-0.79]	0.77 [0.72-0.81]	0.78 [0.77-0.80]	0.78 [0.76-0.80]	0.79 [0.74-0.83]
Zhejiang	0.82 [0.79-0.85]	0.80 [0.76-0.83]	0.79 [0.73-0.84]	0.83 [0.80-0.86]	0.83 [0.81-0.86]	0.81 [0.75-0.86]
Hunan	0.81 [0.79-0.84]	0.79 [0.77-0.82]	0.83 [0.78-0.86]	0.76 [0.74-0.79]	0.76 [0.74-0.79]	0.79 [0.74-0.83]

Table B.3: Discrimination and calibration performance for total stroke and stroke pathological types in men and women (age 55-84) using 2017 Framingham Stroke Risk Profile (FSRP).

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

Note: AF was not recorded in CKB.

Table B.4: Discrimination and calibration performance for total stroke and stroke pathological types in men and women (age 55-84) using 1991 Framingham Stroke Risk Equation (Wolf) [50].

	Men		Women	
	Discrimination AUC [95% CI]	Calibration χ^2	Discrimination AUC [95% CI]	Calibration χ^2
Total Stroke				
2017 FSRP	0.68 [0.66-0.69]	680	0.67 [0.66-0.68]	1,984
+ Recalibration	0.68 [0.66-0.69]	38	0.67 [0.66-0.68]	104
+ Refitting	0.70 [0.69-0.71]	36	0.68 [0.67-0.69]	23
+ Area stratification	0.73 [0.72-0.74]	28	0.74 [0.73-0.75]	53
+ Additional risk factors	0.74 [0.73-0.75]	34	0.75 [0.74-0.76]	50
Ischaemic Stroke (IS)				
2017 FSRP	0.67 [0.65-0.68]	373	0.66 [0.64-0.67]	1,537
+ Recalibration	0.67 [0.65-0.68]	39	0.66 [0.64-0.67]	144
+ Refitting	0.69 [0.68-0.71]	38	0.67 [0.66-0.68]	26
+ Area stratification	0.75 [0.74-0.76]	25	0.75 [0.74-0.76]	21
+ Additional risk factors	0.76 [0.74-0.77]	26	0.75 [0.74-0.76]	22
Haemorrhagic Stroke (HS)				
2017 FSRP	0.71 [0.68-0.73]	742	0.70 [0.67-0.72]	31
+ Recalibration	0.71 [0.68-0.73]	18	0.70 [0.67-0.72]	28
+ Refitting	0.72 [0.70-0.74]	9	0.71 [0.69-0.73]	12
+ Area stratification	0.75 [0.72-0.77]	4	0.75 [0.73-0.77]	23
+ Additional risk factors	0.76 [0.73-0.78]	3	0.77 [0.74-0.79]	11

Note: AF was not recorded in CKB.

Table B.5: Discrimination and calibration performance for total stroke and stroke pathological types in men and women (age 55-84) using 1994 Framingham Stroke Risk Equation (D'Agostino) [51].

	Men		Women	
	Discrimination AUC [95% CI]	Calibration χ^2	Discrimination AUC [95% CI]	Calibration χ^2
Total Stroke				
2017 FSRP	0.67 [0.66-0.69]	N/A	0.67 [0.66-0.68]	N/A
+ Recalibration	0.67 [0.66-0.69]	43	0.67 [0.66-0.68]	101
+ Refitting	0.70 [0.69-0.71]	38	0.68 [0.67-0.69]	26
+ Area stratification	0.73 [0.72-0.74]	27	0.74 [0.73-0.75]	54
+ Additional risk factors	0.74 [0.73-0.75]	34	0.75 [0.74-0.76]	50
Ischaemic Stroke (IS)				
2017 FSRP	0.66 [0.65-0.68]	N/A	0.66 [0.65-0.67]	N/A
+ Recalibration	0.66 [0.65-0.68]	42	0.66 [0.65-0.67]	139
+ Refitting	0.69 [0.68-0.71]	39	0.67 [0.66-0.68]	31
+ Area stratification	0.75 [0.73-0.76]	22	0.75 [0.73-0.76]	25
+ Additional risk factors	0.76 [0.74-0.77]	26	0.75 [0.74-0.76]	22
Haemorrhagic Stroke (HS)				
2017 FSRP	0.71 [0.68-0.73]	N/A	0.70 [0.68-0.73]	N/A
+ Recalibration	0.71 [0.68-0.73]	13	0.70 [0.68-0.73]	29
+ Refitting	0.72 [0.70-0.74]	10	0.71 [0.69-0.74]	13
+ Area stratification	0.74 [0.72-0.77]	8	0.76 [0.73-0.78]	53
+ Additional risk factors	0.76 [0.73-0.78]	3	0.77 [0.74-0.79]	11

Note: AF was not recorded in CKB.

N/A: Calibration could not be assessed for original D'Agostino models because baseline survival curves were not reported.

Table B.6: Comparison of performance of different models for prediction of total stroke and stroke pathological types in men and women (with modified ordering of incremental changes compared to Table 4.2).

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

Note: AF was not recorded in CKB. Shaded rows indicate changes relative to Table 4.2 due to including additional risk factors prior to area stratification.



Chapter 5 Supplementary Materials

C.1 Details of Cox and ML Model Development

C.1.1 Cox Models

Cox models were derived separately for men and women with stratification by 10 geographic areas. For each model, separate baseline survival functions were developed for each CKB area (10 in total). The baseline hazard function (and corresponding baseline survival function) were derived from the non-parametric Breslow estimator [84, 85]. 10-fold cross-validated LASSO regularisation was used (within the training set) for selecting a subset of risk factors from all CKB variables. LASSO regularisation was performed separately for each model, yielding slightly different numbers of selected risk factors (66 risk factors for men, 70 risk factors for women). The specifics of the variable selection process have been previously described [86, 87]. The final Cox models for men and women were used for predictions at all time scales (within 9 years, 0-3 years, 3-6 years, and 6-9 years from baseline).

C.1.2 Random Survival Forest (RSF) Models

RSF models were derived separately for men and women using the `ranger()` function of the `ranger` package version 0.12.1 [114].

For men, number of trees was fixed at 150, and a grid search was performed with `mtry` values of [5, 15, 25, 40, 50] and `min.node.size` values of [10, 100, 1000] within the training set. Best performance in the training set was achieved with 150 trees, `mtry`=40, and `min.node.size`=100.

For women, number of trees was fixed at 150, and a grid search was performed with `mtry` values of [5, 15, 25, 40, 50] and `min.node.size` values of [10, 100, 1000] within the training set. Best performance in the training set was achieved with 150 trees, `mtry`=40, and `min.node.size`=100.

The final RSF models for men and women were used for predictions at all time scales (within 9 years, 0-3 years, 3-6 years, and 6-9 years from baseline).

C.1.3 Logistic Regression (LR) Models

LR models were derived separately for both men and women using the `LogisticRegressionCV()` function of the scikit-learn toolkit version 0.19.2 [96]. For both men and women, L1 penalty was applied with the liblinear solver, and regularisation strength ($10^{-4} < C < 10^4$) was optimized using k-fold cross validation in the training set, (k=5 folds selected based on runtime considerations). Regularisation strength was re-tuned, and new LR models derived, for each time scale (within 9 years, 0-3 years, 3-6 years, and 6-9 years from baseline).

C.1.4 Support Vector Machine (SVM) Models

SVM models were derived separately for both men and women using the `LinearSVC()` function of the scikit-learn toolkit version 0.19.2 [96]. For both men and women, L2 penalty was applied and squared hinge loss was specified. The regularisation parameter ($0.1 \leq C \leq 10$) was tuned using k-fold cross validation in the training set (k=3 folds selected based on runtime considerations). To calibrate the SVM decision function to meaningful probabilities, isotonic regression was implemented with k-fold cross-validation in the training set (k=2 folds selected based on runtime

considerations). Regularisation strength was re-tuned, isotonic regression re-implemented, and new SVM models derived, for each time scale (within 9 years, 0-3 years, 3-6 years, and 6-9 years from baseline).

C.1.5 Gradient Boosted Tree (GBT) Models

The `GradientBoostingClassifier()` function of the scikit-learn toolkit version 0.19.2 [96] was used to derive GBT models separately for both men and women.

For men, `learning_rate` was initially fixed at 0.2, `min_samples_split` fixed at 1000, `min_samples_leaf` fixed at 100, `max_depth` fixed at 8, `max_features` set to "sqrt", and `subsample` fixed at 0.8. A first grid search was performed, with k-fold cross validation in the training set (k=3 folds selected based on runtime considerations), to tune `n_estimators` with values ranging from 20 to 100. After setting `n_estimators`, a second grid search was performed, with 3-fold cross validation, to tune `max_depth` (values ranging from 1 to 10) and `min_samples_split` (values ranging from 500 to 1500). A third grid search was performed, with 3-fold cross validation, to tune `min_samples_leaf` (values ranging from 1 to 6). A fourth grid search was performed, with 3-fold cross validation, to tune `max_features` (values ranging from 10 to 60). A fifth grid search was performed, with 3-fold cross validation, to tune `subsample` (values ranging from 0.6 to 0.9). A sixth grid search was performed, with 3-fold cross validation, to tune `learning_rate` (values ranging from 0.01 to 0.2) and `n_estimators` (values ranging from 60 to 1200). The final tuned model for men had the following hyperparameters: `learning_rate` = 0.02, `n_estimators` = 1200, `max_depth` = 5, `min_samples_split` = 900, `min_samples_leaf` = 1, `subsample` = 0.9, and `max_features` = 25.

For women, an identical hyperparameter methodology was employed. The final tuned model for women had the following hyperparameters: `learning_rate` = 0.01, `n_estimators` = 1200, `max_depth` = 7, `min_samples_split` = 1100, `min_samples_leaf` = 1, `subsample` = 0.9, and `max_features` = 30.

The final tuned hyperparameters were kept the same for predictions at each time scale (within 9 years, 0-3 years, 3-6 years, and 6-9 years from baseline), but the GBT models for men and women were retrained for each prediction task.

C.1.6 Multilayer Perceptron (MLP) Models

MLP models were derived separately for both men and women using keras version 2.3.1 [115].

For men, an architecture was first selected. To perform a binary classification task with 143 inputs, the input layer of the MLP was set at 143 nodes and the output layer set at 1 node with a sigmoid activation function. To select the number of hidden layers and the number of nodes per hidden layer, a grid search was performed with k-fold cross-validation (k=3 folds selected based on runtime considerations) over 1, 3, 5, or 10 hidden layers and 5, 10, 25, 50, 75, or 100 nodes per hidden layer. Since each model had different capacity and would reach a different level of fitting / overfitting for the same number of epochs, we used early stopping (with a 20 epoch maximum and a patience of 5 epochs) to allow all models to train until binary cross-entropy loss in the validation set was sufficiently minimised. For these tests, we used the stochastic gradient descent optimiser with a default learning rate of 0.01, a default batch size of 32, a default "glorot_uniform" initialisation of kernel weights, and a default initialisation of zeros for the bias weights.

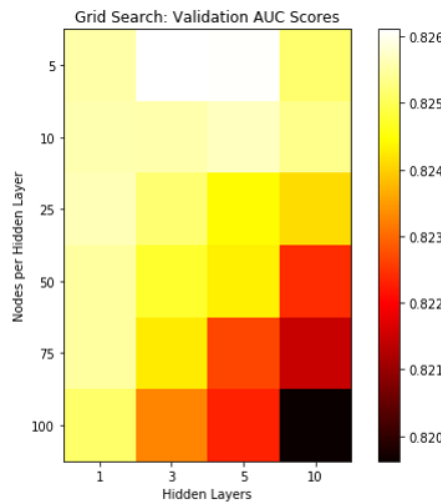


Figure C.1: Grid search for MLP architecture selection.

The results of the grid search for architecture selection are shown in Figure C.1. The best performing model (as assessed by validation AUROC) had 3 hidden layers with 5 nodes each, and the worst performing model had 10 hidden layers with 100 nodes per layer. Figure C.2 shows the model accuracy and loss in the training and validation sets for both the best-performing and worst-performing models. It was observed that 20 epochs provided sufficient training for the validation loss to decrease and converge in the best-performing model, while in the worst-performing model, the model failed to improve at all due to its over-complexity.

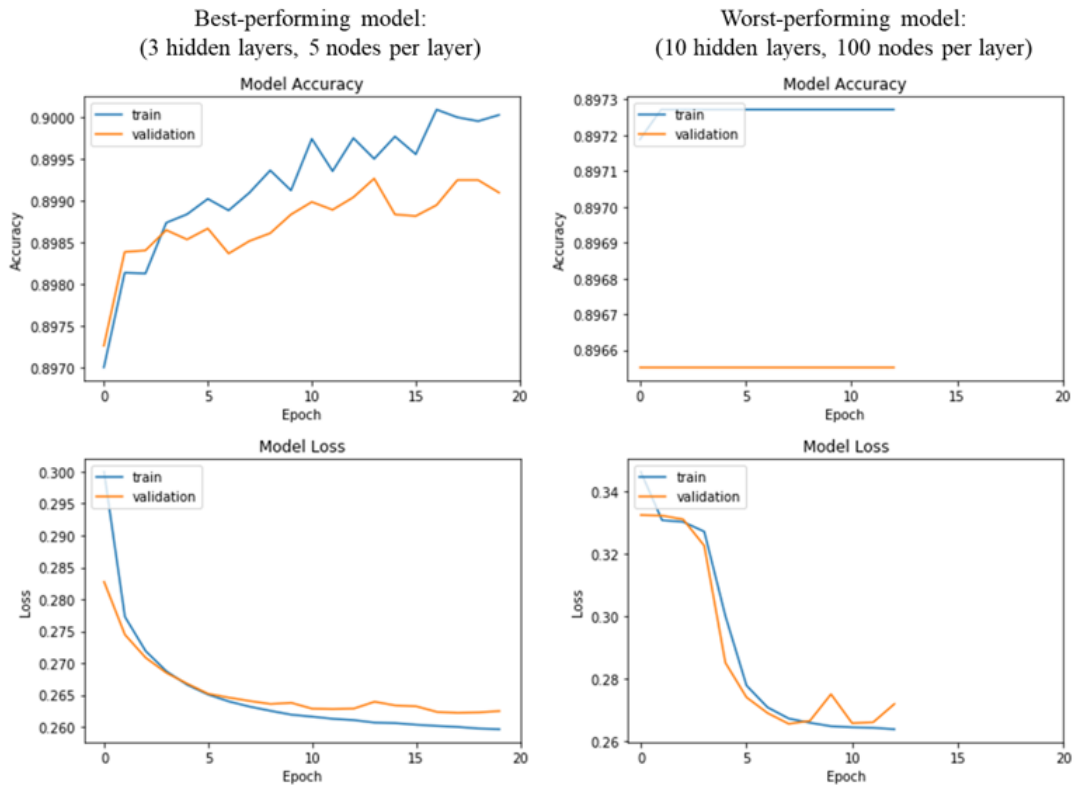


Figure C.2: Model loss and accuracy against epochs for best-performing and worst-performing MLP model architectures.

To select the best activation function for the layers of the MLP (with the exception of the output layer), a 3-fold cross-validated comparison was performed to select between ReLU, tanh, and sigmoid activation functions in the model architecture selected above. ReLU activation functions provided the best performance with a mean validation AUROC of 0.826.

Next, to find appropriate the appropriate learning rate and batch size for training the MLP, another 3-fold cross-validated grid search was performed over learning rates of 0.001, 0.01, or 0.1 and batch sizes of 32, 128, or 512. For these tests, the maximum number of epochs was increased from 20 to 50 to allow for additional training time for models with small learning rates. The best-performing model (as assessed by validation AUROC) used a learning rate of 0.1 and a batch size of 512.

To test if batch normalisation would improve performance, a 3-fold cross-validated comparison was performed to select between the tuned MLP without batch normalisation, with batch normalisation implemented before the non-linear tanh activation function, and with batch normalisation implemented after the non-linearity. A default momentum of 0.99 was used. However, batch normalisation did not have any effect on the validation AUROC.

Table C.1: Optimal parameter values and validation AUROCs for various MLP regularisation techniques.

	Early Stop	Dropout	L1	L2
Best Parameter Value	N/A	0.1	0.0001	0.001
Mean Validation AUROC (SD)	0.8260 (0.0021)	0.8174 (0.0018)	0.8264 (0.0013)	0.8261 (0.0016)

Up to this point in the hyperparameter tuning process, early stopping was used to prevent overfitting of the MLPs. However, overfitting could also be prevented with dropout, L1, or L2 regularisation. Consequently, early stopping was turned off and compared with these alternative regularisation techniques using a 3-fold cross-validated approach. Dropout was tested with dropout ratios of 0.1, 0.2, 0.3, 0.4, and 0.5, with a ratio of 0.1 yielding the best validation AUROC. L1 and L2 regularisation were both tested with regularisation factors of 0.0001, 0.001, 0.01, and 0.1. A regularisation factor of 0.0001 yielded the highest validation AUROC for L1 regularisation, and a regularisation factor of 0.001 yielded the highest validation AUROC for L2 regularization. As shown in the results summary of Table C.1, the weakly L1-regularised model (with regularisation factor of 0.0001) yielded the highest mean validation AUROC of 0.8264.

Using this tuning process, the final MLP for men had the following hyperparameters: SGD optimiser, "glorot_uniform" initialisation of kernel weights, zero initialisation for bias weights, a 143 node input layer, 3 hidden layers with 5 nodes each and ReLU activation functions, a 1 node output layer with a sigmoid activation function, learning rate = 0.1, batch size = 512, and L1 regularisation with a regularisation factor of 0.0001.

For women, a similar hyperparameter tuning process was implemented, resulting in a final MLP with the following hyperparameters: SGD optimiser, "glorot_uniform" initialisation of kernel weights, zero initialisation for bias weights, a 143 node input layer, 3 hidden layers with 10 nodes each and tanh activation functions, a 1 node output layer with a sigmoid activation function, learning rate = 0.01, batch size = 32, and L2 regularisation with a regularisation factor of 0.0001.

The final tuned hyperparameters were kept the same for predictions at each time scale (within 9 years, 0-3 years, 3-6 years, and 6-9 years from baseline), but the MLP models for men and women were retrained for each prediction task.

C.1.7 Decision Tree for Selection between Cox and GBT

Decision tree models for selection between Cox and GBT were derived separately for both men and women using the `DecisionTreeClassifier()` function of the scikit-learn toolkit version 0.19.2 [96].

For men, a grid search was performed, with k-fold cross validation (k=3 folds selected based on runtime considerations), to tune `max_depth` (values ranging from 1 to 10), `min_sample_split` (values ranging from 2 to 10), and `max_features` (values ranging from 10 to 60). The final tuned model for men had a mean validation AUROC of 0.707 and had the following hyperparameters: `max_depth` = 6, `max_features` = 45, `min_samples_split` = 7. All other hyperparameters were set to default values.

For women, an identical hyperparameter methodology was employed. The final tuned model for men had a mean validation AUROC of 0.728 and had the following

hyperparameters: `max_depth = 6`, `max_features = 55`, `min_samples_split = 8`. All other hyperparameters were set to default values.

C.2 Additional Figures and Tables

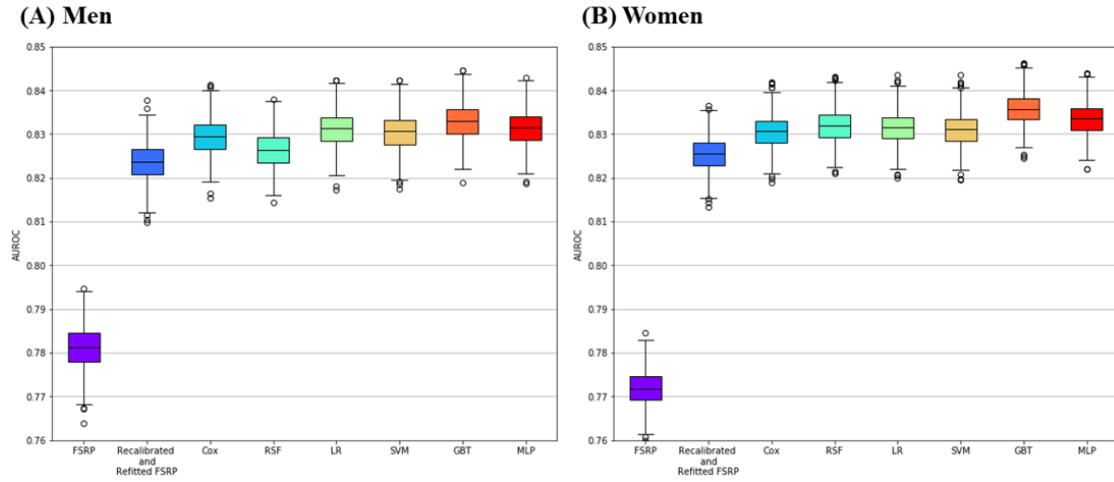


Figure C.3: AUROCs for prediction of 9 year risk of stroke (1000 bootstrapped samples from the validation set).

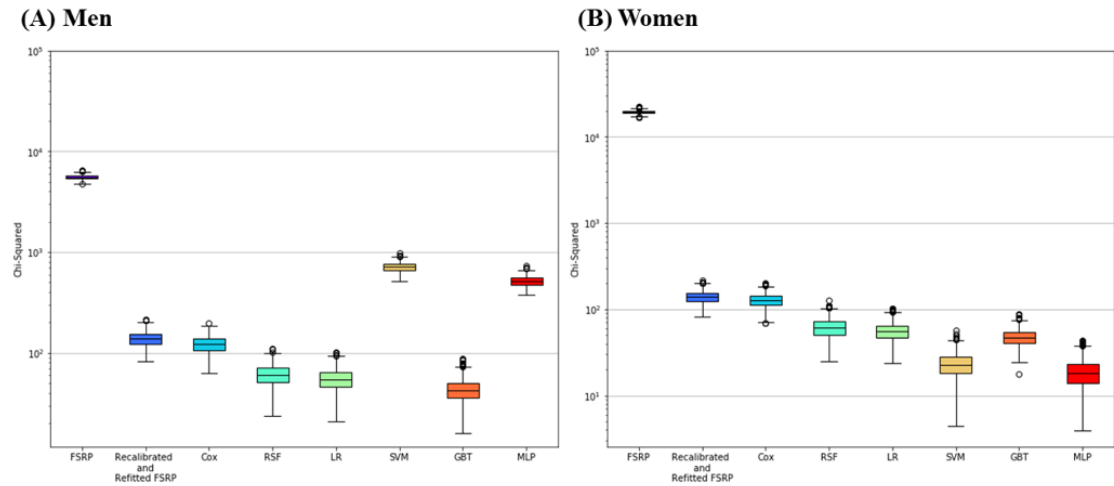


Figure C.4: Hosmer-Lemeshow/Nam-D'Agostino test statistics for prediction of 9 year risk of stroke (1000 bootstrapped samples from test set).

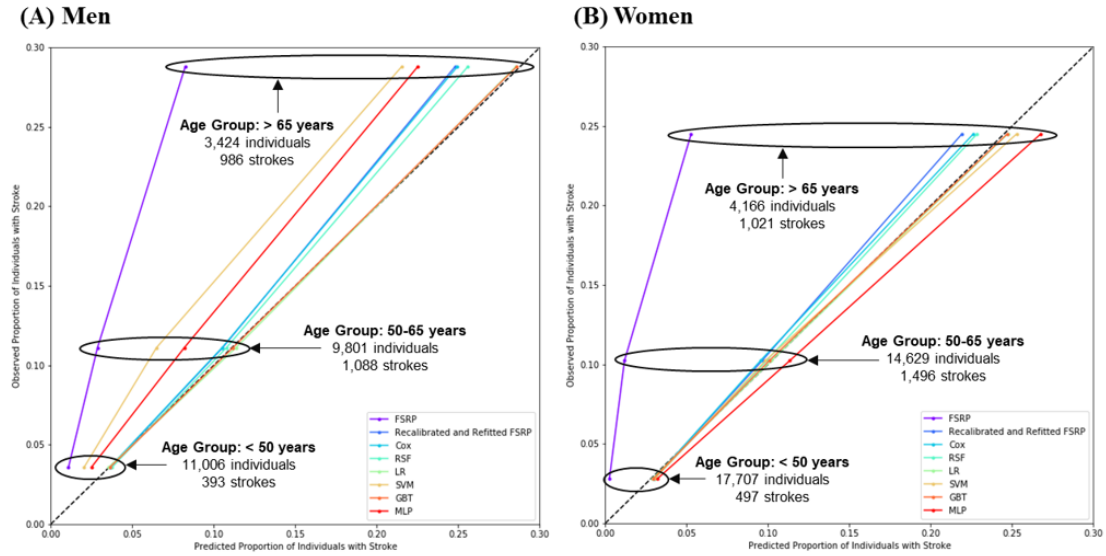


Figure C.5: Calibration plots by age group for the 2017 Framingham Stroke Risk Profile (FSRP), a recalibrated and refitted FSRP, Cox, random survival forest (RSF), logistic regression (LR), support vector machine (SVM), gradient boosted tree (GBT), and multilayer perceptron (MLP) models in (A) men and (B) women. Each point represents an age group in the CKB cohort.

Table C.2: Discrimination and calibration performance of Cox and ML models for risk prediction of stroke from 0-3 years.

Model Type	Men		Women	
	Discrimination AUROC [95% CI]	Calibration χ^2 [95% CI]	Discrimination AUROC [95% CI]	Calibration χ^2 [95% CI]
FSRP	0.800 [0.783-0.816]	1,191 [956-1,443]	0.803 [0.787-0.817]	5,146 [4,278-6,095]
Recalibrated and Refitted FSRP	0.832 [0.818-0.846]	51 [29-80]	0.834 [0.820-0.847]	70 [42-103]
Cox	0.840 [0.826-0.854]	53 [33-78]	0.840 [0.826-0.852]	62 [37-94]
RSF	0.835 [0.821-0.848]	26 [14-40]	0.846 [0.833-0.859]	38 [22-57]
LR	0.867 [0.854-0.879]	20 [9-33]	0.864 [0.850-0.876]	29 [14-52]
SVM	0.864 [0.850-0.877]	309 [217-433]	0.862 [0.848-0.875]	20 [9-36]
GBT	0.864 [0.850-0.877]	36 [20-59]	0.867 [0.854-0.879]	36 [19-60]
MLP	0.864 [0.851-0.877]	38 [17-69]	0.865 [0.852-0.877]	29 [13-54]

Table C.3: Discrimination and calibration performance of Cox and ML models for risk prediction of stroke from 3-6 years.

Model Type	Men		Women	
	Discrimination	Calibration	Discrimination	Calibration
	AUROC [95% CI]	χ^2 [95% CI]	AUROC [95% CI]	χ^2 [95% CI]
FSRP	0.768 [0.752-0.783]	3,039 [2,593-3,608]	0.764 [0.751-0.777]	7,095 [6,135-8,147]
Recalibrated and Refitted FSRP	0.810 [0.797-0.824]	105 [65-152]	0.817 [0.805-0.829]	70 [42-103]
Cox	0.815 [0.802-0.829]	75 [42-118]	0.822 [0.809-0.833]	44 [23-70]
RSF	0.814 [0.800-0.828]	44 [23-70]	0.823 [0.811-0.836]	39 [21-61]
LR	0.837 [0.824-0.850]	20 [8-37]	0.840 [0.829-0.852]	23 [12-38]
SVM	0.836 [0.824-0.849]	353 [227-545]	0.839 [0.828-0.850]	22 [9-39]
GBT	0.839 [0.826-0.851]	30 [13-54]	0.844 [0.833-0.855]	24 [12-41]
MLP	0.837 [0.824-0.885]	386 [289-497]	0.842 [0.831-0.853]	27 [10-50]

Table C.4: Discrimination and calibration performance of Cox and ML models for risk prediction of stroke from 6-9 years.

Model Type	Men		Women	
	Discrimination	Calibration	Discrimination	Calibration
	AUROC [95% CI]	χ^2 [95% CI]	AUROC [95% CI]	χ^2 [95% CI]
FSRP	0.735 [0.720-0.748]	2,538 [2,141-2,967]	0.729 [0.717-0.741]	9,655 [8,505-10,860]
Recalibrated and Refitted FSRP	0.778 [0.766-0.790]	122 [76-177]	0.790 [0.780-0.801]	128 [82-182]
Cox	0.781 [0.769-0.793]	146 [93-208]	0.794 [0.782-0.804]	134 [85-189]
RSF	0.775 [0.762-0.787]	64 [35-102]	0.794 [0.784-0.805]	65 [36-100]
LR	0.805 [0.792-0.817]	44 [23-69]	0.810 [0.800-0.821]	63 [37-94]
SVM	0.805 [0.792-0.817]	227 [157-302]	0.809 [0.800-0.820]	23 [10-41]
GBT	0.804 [0.791-0.817]	39 [19-64]	0.815 [0.804-0.825]	46 [25-71]
MLP	0.804 [0.792-0.816]	152 [97-214]	0.811 [0.801-0.821]	36 [20-56]

Table C.5: Confusion matrices for Cox-only, GBT-only, and ensemble approaches for screening of high-risk individuals in the test set.

Men			Women		
Cox-Only Approach			Cox-Only Approach		
	Stroke	No Stroke		Stroke	No Stroke
Identified High Risk	321	954	Identified High Risk	359	1,161
Identified Low Risk	102	2,904	Identified Low Risk	166	4,755
GBT-Only Approach			GBT-Only Approach		
	Stroke	No Stroke		Stroke	No Stroke
Identified High Risk	338	992	Identified High Risk	388	1,317
Identified Low Risk	85	2,866	Identified Low Risk	137	4,599
Ensemble Approach			Ensemble Approach		
	Stroke	No Stroke		Stroke	No Stroke
Identified High Risk	322	919	Identified High Risk	354	1,136
Identified Low Risk	101	2,939	Identified Low Risk	171	4,780

D

Chapter 6 Supplementary Materials

D.1 Hyperparameter Tuning of GBT Models

The `GradientBoostingClassifier()` function of the scikit-learn toolkit version 0.19.2 [96] was used to derive all GBT models separately for both men and women. All hyperparameters were tuned to maximise AUC in the validation set.

For each GBT model, `learning_rate` was initially fixed at 0.1, `min_samples_split` fixed at 2, `min_samples_leaf` fixed at 1, `max_depth` fixed at 8, `max_features` set to "sqrt", and `subsample` fixed at 0.8. A first grid search was performed, with k-fold cross validation in the training set (k=3 folds selected based on runtime considerations), to tune `n_estimators` with values ranging from 5 to 100. After setting `n_estimators`, a second grid search was performed, with 3-fold cross validation, to tune `max_depth` (values ranging from 1 to 10) and `min_samples_split` (values ranging from 2 to 102). A third grid search was performed, with 3-fold cross validation, to tune `min_samples_leaf` (values ranging from 1 to 10). A fourth grid search was performed, with 3-fold cross validation, to tune `max_features` (values ranging from 1 to 406). A fifth grid search was performed, with 3-fold cross validation, to tune `subsample` (values ranging from 0.6 to 0.9). A sixth grid search was performed, with 3-fold cross validation, to tune `learning_rate` (values ranging from 0.001 to 0.1) and `n_estimators` (values

ranging from 50 to 5000). Finally, after all hyperparameters were selected, 3-fold cross validation was used to calibrate (using isotonic regression) and train the final model. The tuned hyperparameters for each GBT model are reported below.

D.1.1 GBT: Single measurement at most recent visit

The final tuned model for men had the following hyperparameters:

`learning_rate = 0.001, n_estimators = 5000, max_depth = 5,`
`min_samples_split = 2, min_samples_leaf = 1, subsample = 0.8, and`
`max_features = 26.`

The final tuned model for women had the following hyperparameters:

`learning_rate = 0.01, n_estimators = 500, max_depth = 5,`
`min_samples_split = 82, min_samples_leaf = 1, subsample = 0.8, and`
`max_features = 26.`

D.1.2 GBT: Sequential measurements at three visits

The final tuned model for men had the following hyperparameters:

`learning_rate = 0.001, n_estimators = 5000, max_depth = 3,`
`min_samples_split = 62, min_samples_leaf = 7, subsample = 0.8, and`
`max_features = 26.`

The final tuned model for women had the following hyperparameters:

`learning_rate = 0.001, n_estimators = 5000, max_depth = 5,`
`min_samples_split = 42, min_samples_leaf = 5, subsample = 0.7, and`
`max_features = 26.`

D.1.3 GBT: Longitudinal summary of sequential measurements at three visits

The final tuned model for men had the following hyperparameters:

`learning_rate = 0.1, n_estimators = 50, max_depth = 3,`
`min_samples_split = 22, min_samples_leaf = 8, subsample = 0.8, and`
`max_features = 126.`

The final tuned model for women had the following hyperparameters:

```
learning_rate = 0.01, n_estimators = 500, max_depth = 3,
min_samples_split = 62, min_samples_leaf = 4, subsample = 0.7, and
max_features = 76.
```

D.1.4 GBT: Longitudinal summary of stroke risk estimates at three visits + Single measurement at most recent visit

The final tuned model for men had the following hyperparameters:

```
learning_rate = 0.001, n_estimators = 50, max_depth = 3,
min_samples_split = 42, min_samples_leaf = 1, subsample = 0.8, and
max_features = 41.
```

The final tuned model for women had the following hyperparameters:

```
learning_rate = 0.001, n_estimators = 500, max_depth = 3,
min_samples_split = 142, min_samples_leaf = 1, subsample = 0.8, and
max_features = 71.
```

D.2 Additional Figures and Tables

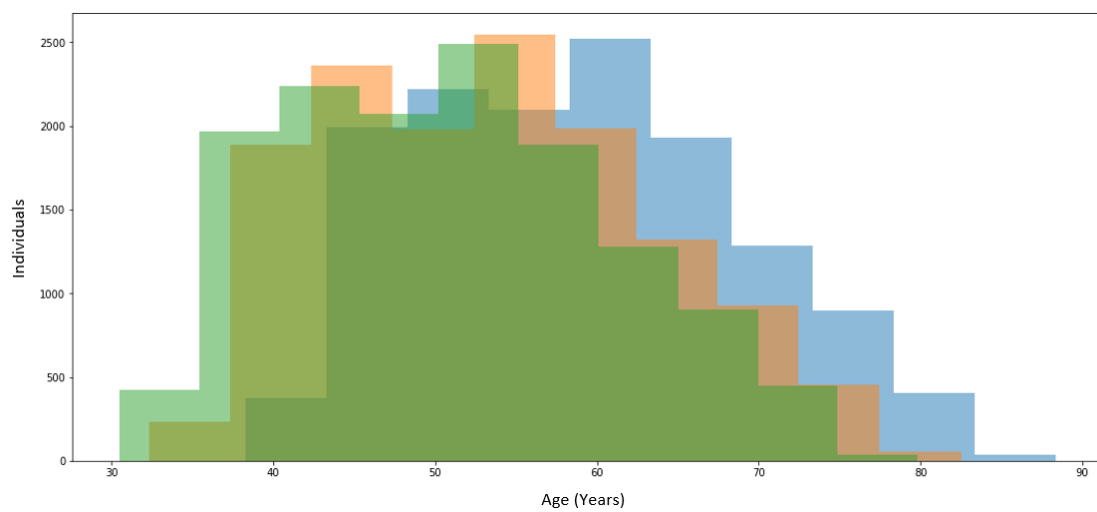


Figure D.1: Distribution of age at baseline (green), resurvey 1 (orange), and resurvey 2 (blue) in included CKB individuals.

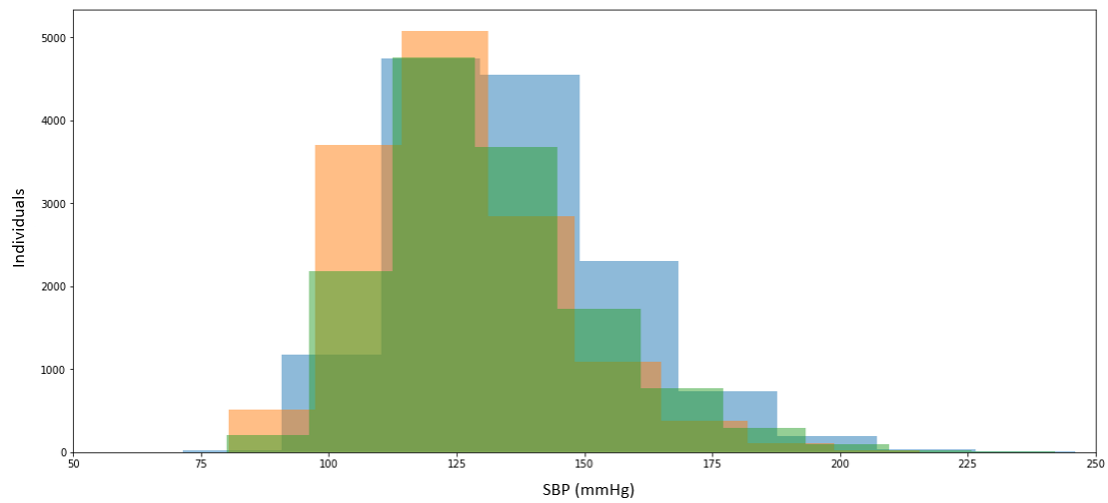


Figure D.2: Distribution of systolic blood pressure (SBP) at baseline (green), resurvey 1 (orange), and resurvey 2 (blue) in included CKB individuals.

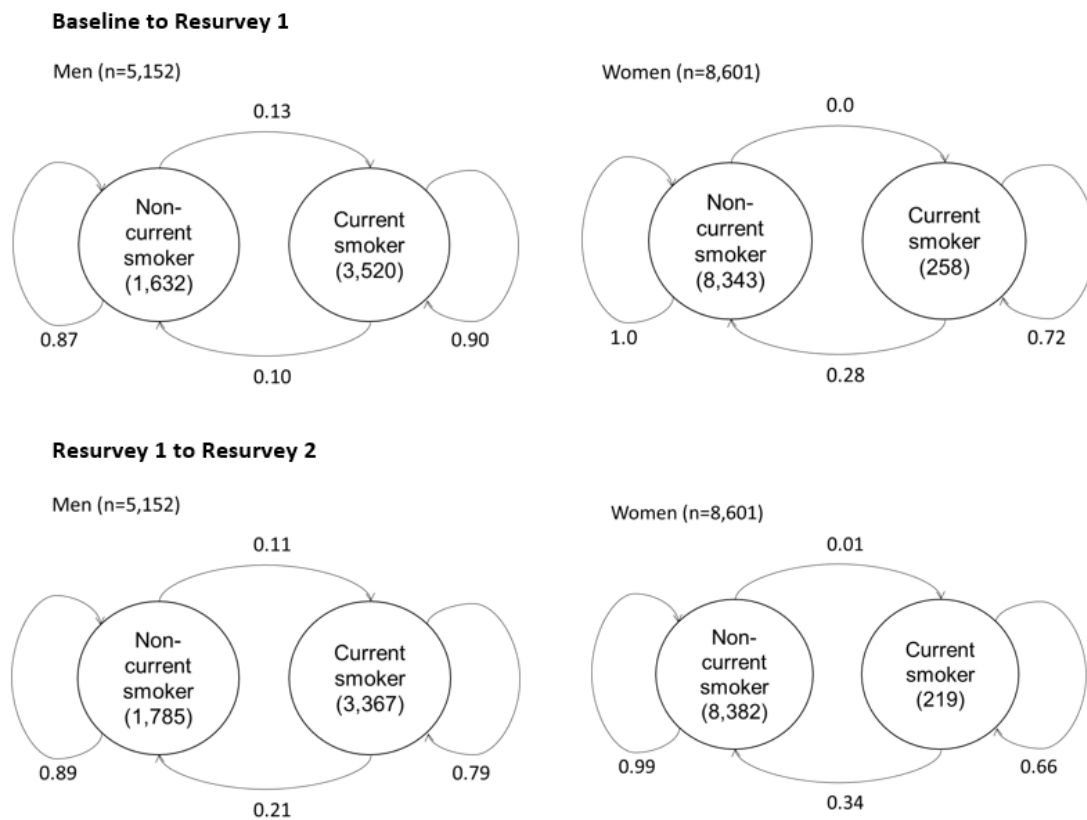


Figure D.3: Distribution of self-reported smoking status and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals.

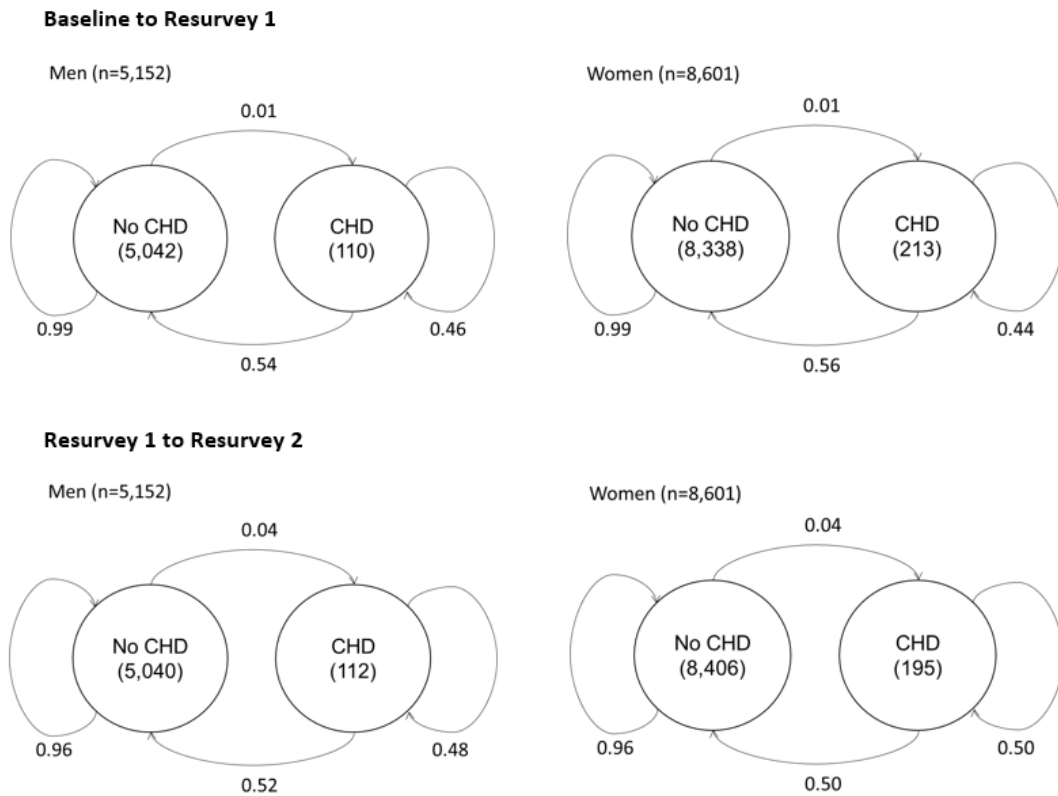


Figure D.4: Distribution of history of coronary heart disease (CHD) and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals. Note: No individuals who self-report a history of a CHD diagnosis should properly report that they have no history of a CHD diagnosis at a subsequent visit. To adjust for these self-reporting errors, in our analyses, we considered an individual to have a history of coronary heart disease if they ever reported CHD at a previous visit.

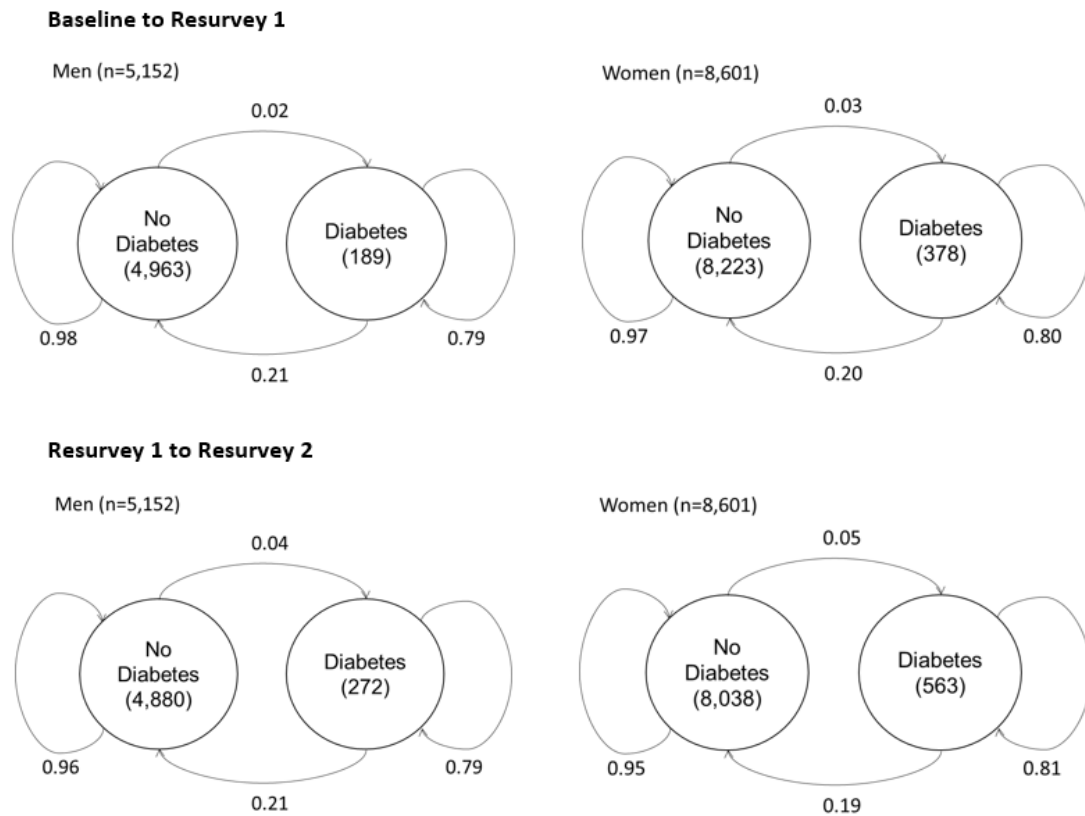


Figure D.5: Distribution of history of diabetes status and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals. Note: No individuals who self-report a history of a diabetes diagnosis or have detected diabetes should properly report that they have no history of a diabetes diagnosis at a subsequent visit. To adjust for these self-reporting errors, in our analyses, we considered an individual to have a history of diabetes if they ever had self-reported or detected diabetes at a previous visit.

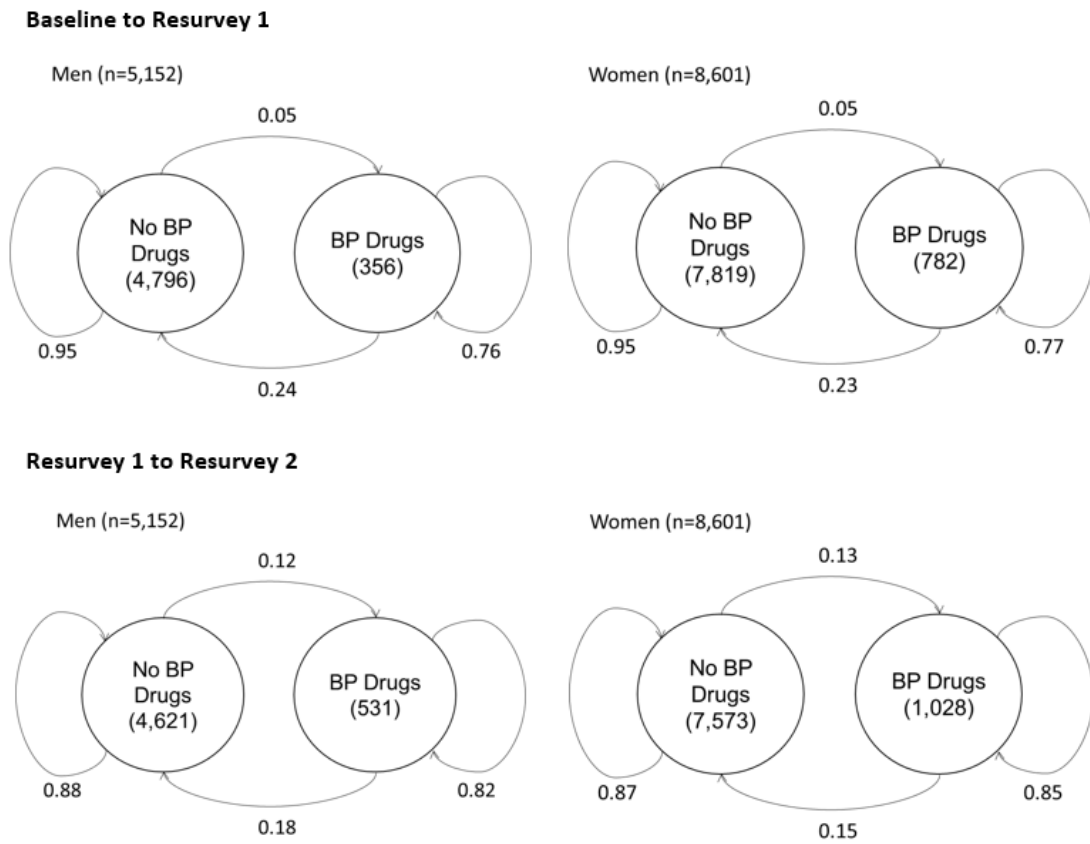


Figure D.6: Distribution of self-reported use of antihypertensive medication and state transitions between baseline, resurvey 1, and resurvey 2 among included CKB individuals.

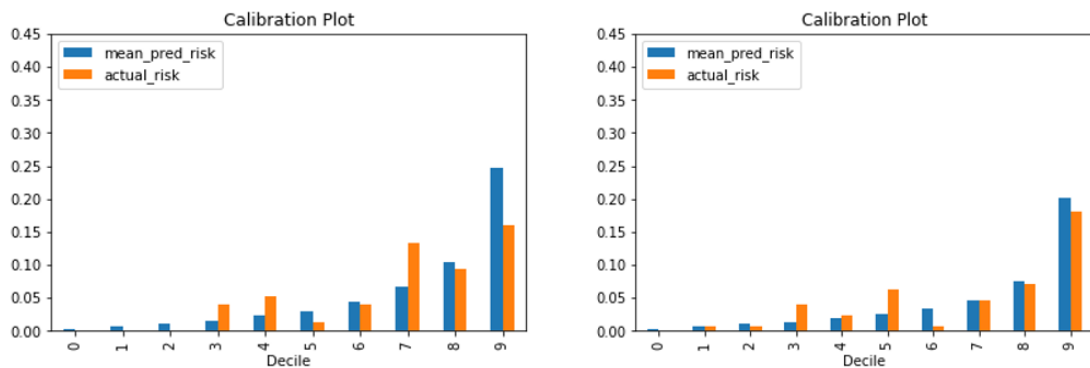


Figure D.7: Calibration plots for “Cox: Single measurement at most recent visit” models in men (left) and women (right).

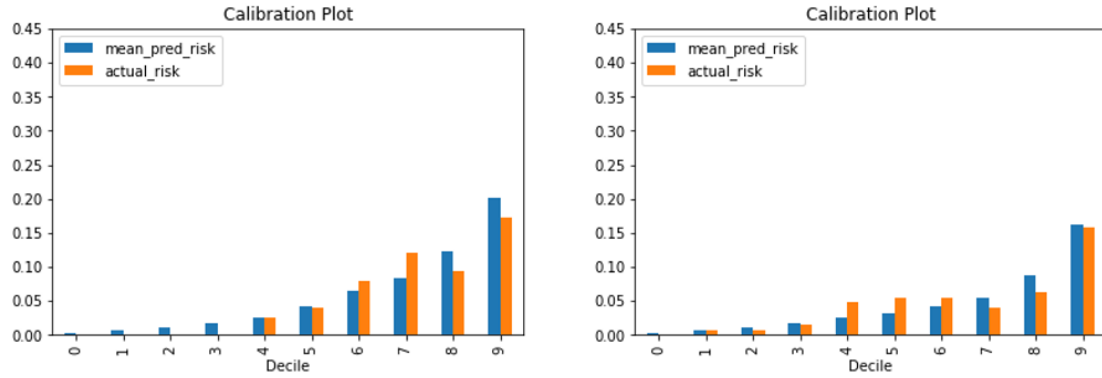


Figure D.8: Calibration plots for “GBT: Single measurement at most recent visit” models in men (left) and women (right).

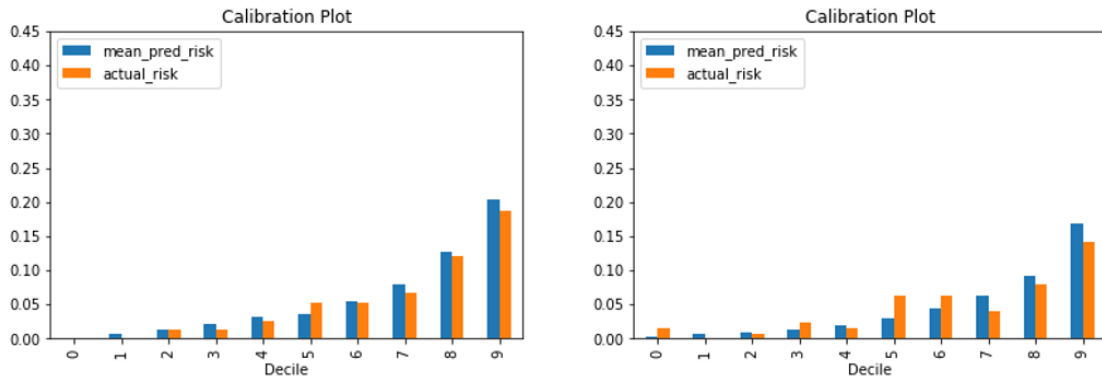


Figure D.9: Calibration plots for “GBT: Sequential measurements at three visits” models in men (left) and women (right).

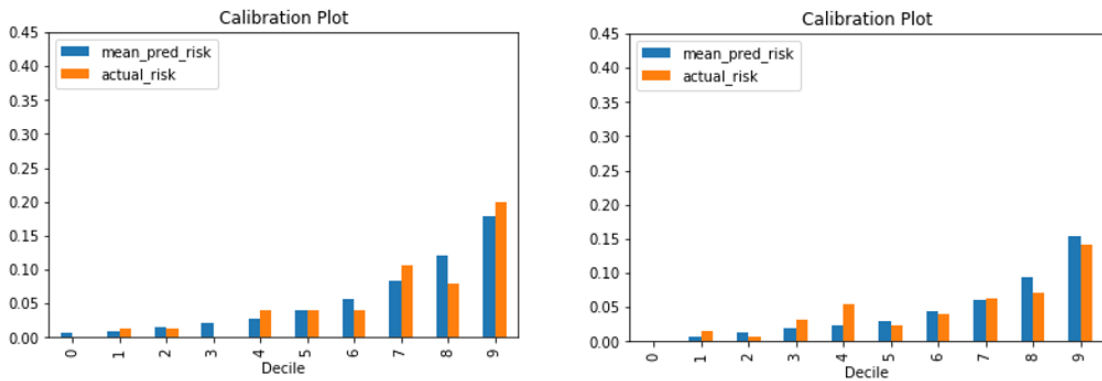


Figure D.10: Calibration plots for “GBT: Longitudinal summary of sequential measurements at three visits” models in men (left) and women (right).

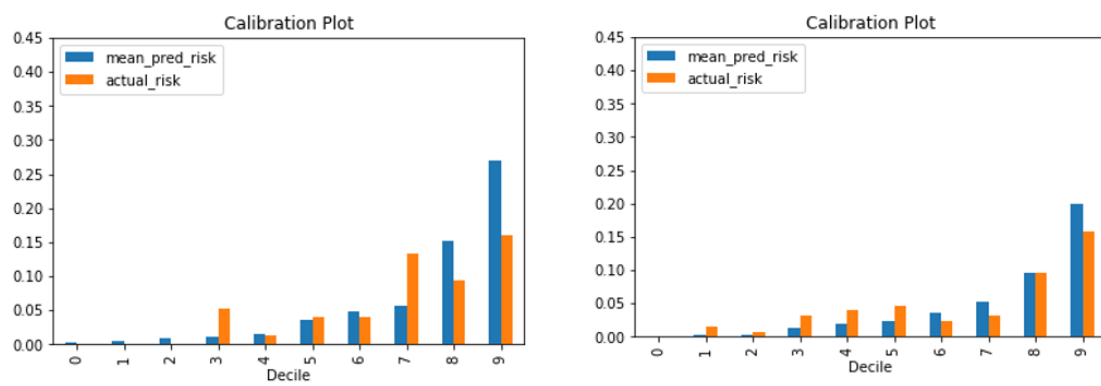


Figure D.11: Calibration plots for “Longitudinal summary of stroke risk estimates at three visits + Single measurement at most recent visit” models in men (left) and women (right).

E

Chapter 7 Supplementary Materials

E.1 Details of GBT Model Development

The `GradientBoostingClassifier()` function of the scikit-learn toolkit version 0.19.2 [96] was used to derive all GBT models separately for both men and women. The GBT models were trained on the subset of individuals in the training set who experienced a first incident stroke (20,186 men; 25,361 women) to predict a probability that individual's first incident stroke would be haemorrhagic conditional on a first stroke event occurring. All hyperparameters were tuned to maximise AUC in the validation set.

For each model, `learning_rate` was initially fixed at 0.1, `min_samples_split` fixed at 1,000, `min_samples_leaf` fixed at 100, `max_depth` fixed at 8, `max_features` set to "sqrt", and `subsample` fixed at 0.8. A first grid search was performed, with k-fold cross validation in the training set (k=3 folds selected based on runtime considerations), to tune `n_estimators` with values ranging from 10 to 100. After setting `n_estimators`, a second grid search was performed, with 3-fold cross validation, to tune `max_depth` (values ranging from 1 to 15) and `min_samples_split` (values ranging from 100 to 1500). A third grid search was performed, with 3-fold cross validation, to tune `min_samples_leaf` (values ranging from 50 to 200). A fourth grid search was performed, with 3-fold cross validation, to

tune `max_features` (values ranging from 5 to 100). A fifth grid search was performed, with 3-fold cross validation, to tune `subsample` (values ranging from 0.6 to 0.9). A sixth grid search was performed, with 3-fold cross validation, to tune `learning_rate` (values ranging from 0.01 to 0.1) and `n_estimators` (values ranging from 50 to 1000).

The final tuned model for men had the following hyperparameters:

`learning_rate = 0.01`, `n_estimators = 700`, `max_depth = 7`, `min_samples_split = 500`, `min_samples_leaf = 120`, `subsample = 0.8`, and `max_features = 30`.

The final tuned model for women had the following hyperparameters:

`learning_rate = 0.01`, `n_estimators = 600`, `max_depth = 11`, `min_samples_split = 1300`, `min_samples_leaf = 90`, `subsample = 0.9`, and `max_features = 25`.

E.2 Additional Figures and Tables

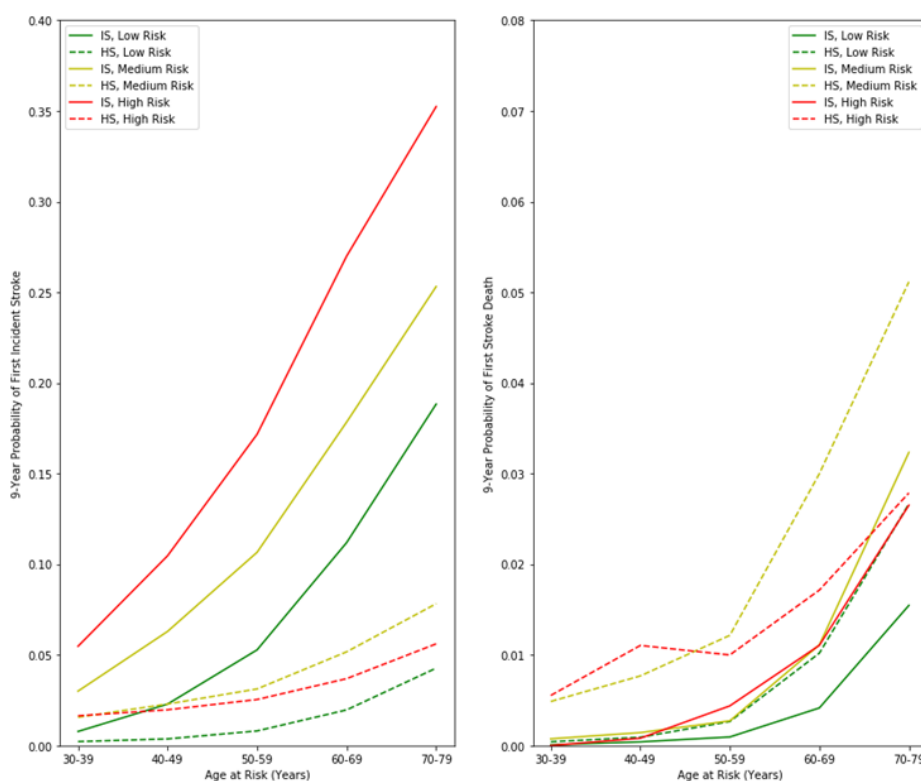
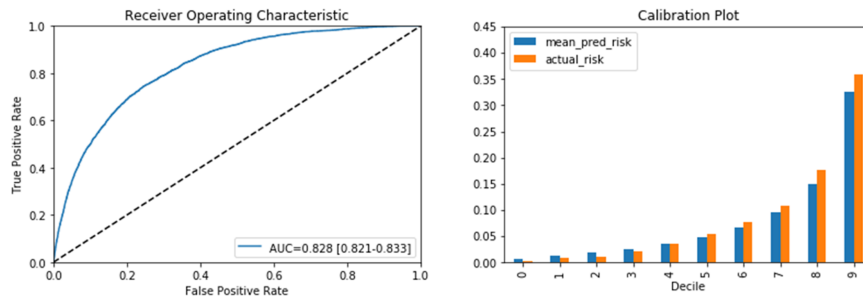


Figure E.1: 9-year incidence and incident mortality from first ischaemic stroke (IS) and first haemorrhagic stroke (HS) cases by age group with stratification by vascular risk.

Men: Ischaemic Stroke



Women: Ischaemic Stroke

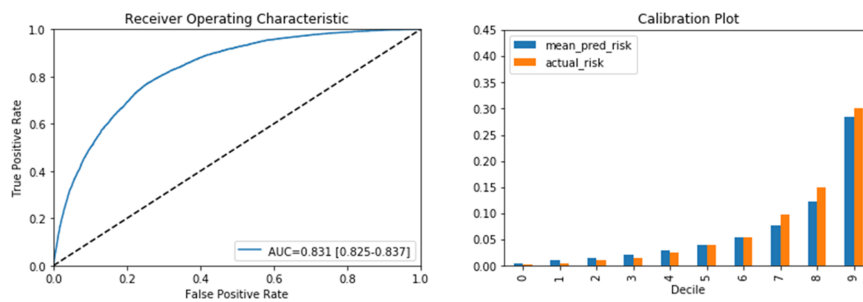
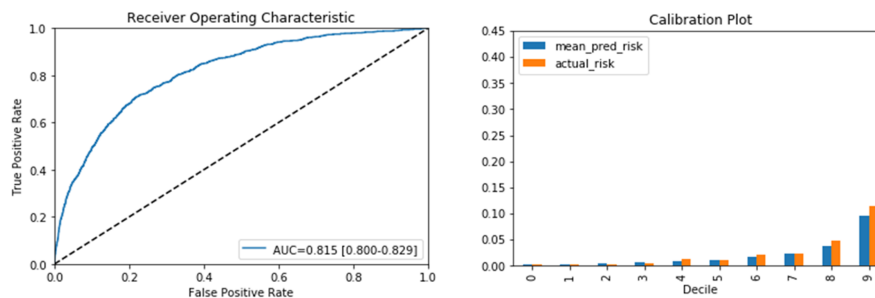


Figure E.2: Discrimination and calibration performance of Cox models for predicting nine-year absolute risks of ischaemic stroke (IS).

Men: Haemorrhagic Stroke



Women: Haemorrhagic Stroke

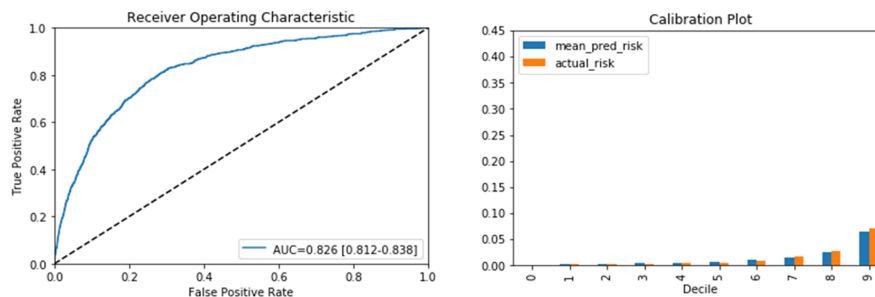


Figure E.3: Discrimination and calibration performance of Cox models for predicting nine-year absolute risks of haemorrhagic stroke (HS).

Table E.1: Baseline characteristics of included participants from the China Kadoorie Biobank with an incident major coronary event (MCE) prior to any stroke event.

	MCE
Number of participants	9,478
<i>Demographic factors</i>	
Age at baseline, mean (s.d.) (years)	62.0 (9.8)
Age 30-39 years, (%)	2.3
Age 40-49 years, (%)	10.9
Age 50-59 years (%)	24.5
Age 60-69 years (%)	37.2
Age 70-79 years (%)	25.1
Female (%)	44.5
Urban (%)	44.9
≥6 years of education (%)	37.9
Household income (≥20,000 yuan per year) (%)	32.2
<i>Medical history and health status</i>	
Diabetes (%)	15.4
Coronary heart disease (%)	13.3
Hypertension (%)	60.1
Blood pressure drugs (%)	23.7
Self-rated poor health status (%)	17.7
<i>Physical measurements, mean (s.d.)</i>	
SBP (mmHg)	143.3 (24.4)
DBP (mmHg)	80.1 (12.7)
BMI (kg m ⁻²)	23.8 (3.9)
<i>Lifestyle factors</i>	
Male ever-regular smokers (%)	78.6
Female ever-regular smokers (%)	8.4
Male ever-regular drinkers (%)	37.1
Female ever-regular drinkers (%)	3.5
Physical activity, mean (s.d.) (hours per day)	4.4 (3.1)
Ever-regular smokers include current- and ex-regular smokers.	
Ever-regular drinkers include current-regular, ex-regular and reduced-intake alcohol drinkers.	

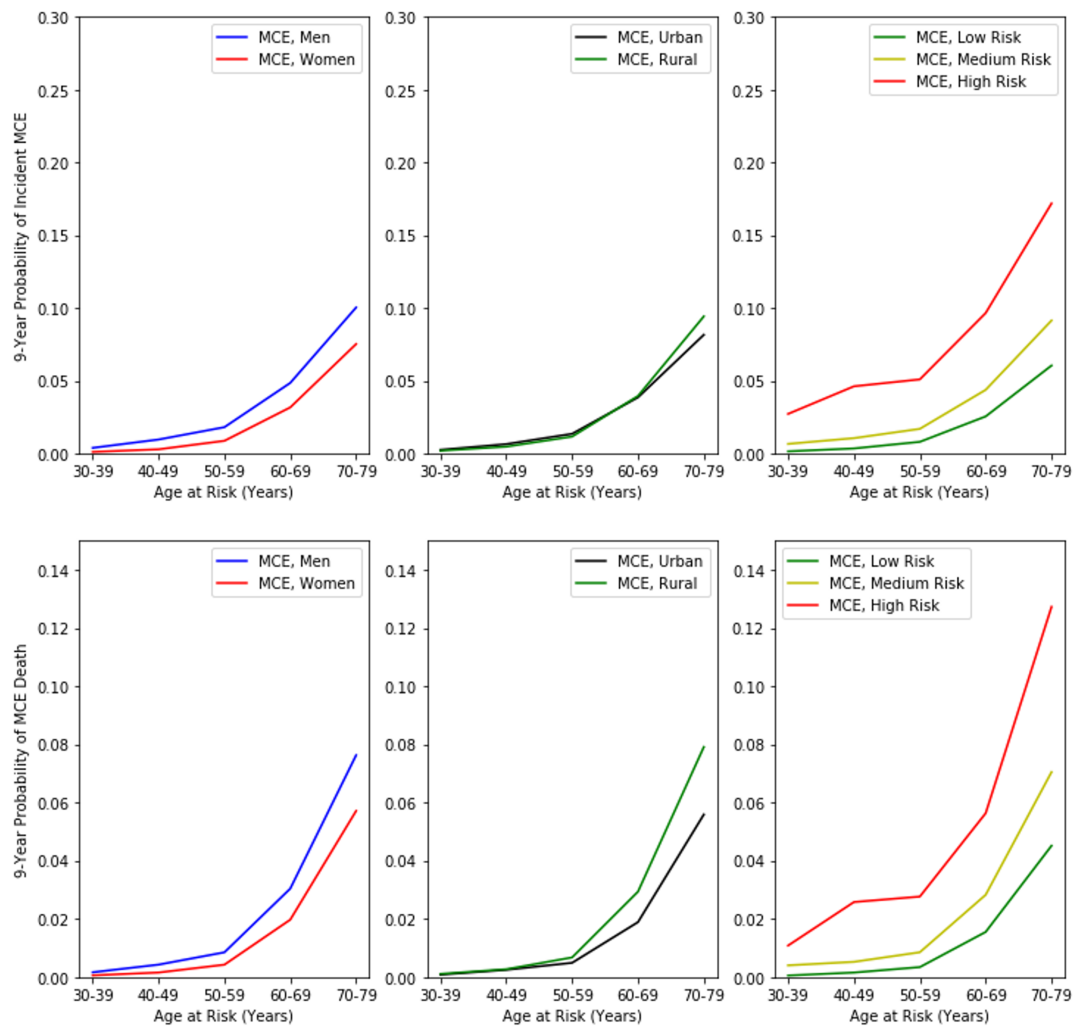


Figure E.4: 9-year probability of incident major coronary event (MCE) and MCE death by age group with stratification by sex, type of area, and vascular risk.

F

Chapter 8 Supplementary Materials

F.1 Implementation of Causative Classification System for Ischemic Stroke in CKB

Primary cases of IS in the China Kadoorie Biobank (CKB) were classified using a web-based implementation of the Causative Classification System for Ischemic Stroke (CCS) [14, 24], which can be accessed at https://ccs.mgh.harvard.edu/ccs_form.php.

In CKB, adjudicators were asked to complete a standardised electronic form, which requested specific details on presenting symptoms and signs and diagnostic criteria, including imaging evidence of haemorrhage or ischaemia, laterality and anatomical location of any cerebral lesions, and additional evidence of extracranial and intracranial arterial disease or sources of embolism. Where necessary, collected variables were slightly adapted to be fitted to the CCS algorithm, as described in further detail below. In addition, CCS has certain limitations for use in East Asian cohorts, which are also discussed below.

F.1.1 Fitting of CKB Variables to CCS

- **Brain Imaging:** In CCS, a lacunar lesion is defined as being <20 mm in diameter, while in CKB, lacunar lesions were defined as being <15 mm in

diameter. CKB also did not include imaging variables for internal watershed infarcts or temporally separate infarcts. While CKB had capacity to identify cases with multiple acute ischaemic lesions in both the left and right anterior or both the anterior and posterior circulations, these were not included for IS subtyping. Finally, in CCS, one criteria for a possible SAO is presentation of “a classical lacunar syndrome in the absence of imaging that is sensitive enough to detect small infarctions.” For the purposes of implementing CCS, MRI was treated as sensitive enough to detect small infarctions, while CT was not.

- **Vascular Imaging:** From collected CKB variables, the relationship of arterial disease to the anatomical site of infarct could not be determined with certainty, which is relevant in CCS for considering parent artery disease. Second, for simplicity, stenosis of $<50\%$ or $\geq 50\%$ was considered if it was determined to be present in any artery (intracranial or extracranial). Third, details of plaque echogenicity in the extracranial arteries only (i.e., hypoechoic or mixed echoic findings) were used as surrogates of thrombus formation and ulceration.
- **Cardiac Investigations and Other Clinical Findings:** First, the CKB variable for systemic embolism, which differentiates “Probable” from “Possible” CE cases was selected in only 26 IS cases and may be underreported. Second, some high- and low- or uncertain risk cardioaortic sources of cerebral embolism considered by CCS were not included in the CKB standardised electronic adjudication form. Third, while prior history of IS, transient monocular blindness, or transient ischaemic attack can be influential in CCS, these were not appropriate in the present CKB analyses, which considered incident primary IS cases only, and excluded participants with prior history of stroke or transient ischaemic attack at baseline. Finally, the CKB standardised electronic adjudication form did not include any means to record “Other causes” of multifocal ischaemic brain injury (e.g., vasculitides, vasculopathies,

and haemostatic or haemodynamic disturbances), which were not considered in the present study.

F.1.2 Limitations of Using CCS in CKB

- **Stroke Risk Estimates:** In CCS, a 2% annual or one-time primary stroke risk threshold was chosen to separate “evident” from “possible” mechanisms. This threshold was chosen because it is the approximate, annual, primary, ipsilateral stroke risk associated with asymptomatic carotid stenosis greater than 50% [24]. However, these estimates are based on Western populations rather than East Asian cohorts and may not be entirely suitable for CKB. In addition, CCS uses an arbitrary cutoff value of 2 for the positive likelihood ratio at which the presence of a feature is considered for distinguishing the most likely aetiological mechanism. This arbitrary cutoff may also be inappropriate for use in East Asian cohorts for similar reasons.

F.2 Details of Multiclass LR Model Development

F.2.1 Data Release

Data used in this study were from CKB version 17.02.

F.2.2 Data Preprocessing

F.2.2.1 Training Set and Test Set Split

Individuals with a recorded primary IS case of evident and probable determined aetiology (n=7,4443) were divided into a training set and test set using a random 85%/15% training/test split and checked to ensure similar proportions of CE/LAA/SAO (Table F.1).

F.2.2.2 Missing Values

For the 7,443 individuals included in the modelling analyses, missing values were very rare among the CKB baseline variables and were imputed in both the training set and test set using the sex-specific means of the non-missing values in the training set.

Table F.1: Proportions of ischaemic stroke (IS) subtypes in the training set and test set.

	Training Set (n=6,326)	Test Set (n=1,117)
	Number of individuals (%)	Number of individuals (%)
CE	136 (2.1%)	24 (2.1%)
LAA	2,036 (32.2%)	359 (32.1%)
SAO	4,154 (65.7%)	734 (65.7%)

As described in the main text, findings of cardioaortic sources of embolism from cardiac tests identified from hospital records (i.e., the “ce_risk” variable in Table 8.1) were also included as inputs to the multiclass LR model. This variable was coded as 0: no identified cardioembolic source (including due to lack of cardiac tests), 1: low- or uncertain risk cardioembolic source, or 2: high-risk cardioembolic source, in accordance with the cardioaortic sources of cerebral embolism recognised by CCS. The “ce_risk” variable did not include any missing values [14, 24].

F.2.3 Multiclass LR Model Construction

A multiclass LR model was derived using the `LogisticRegressionCV()` function of the scikit-learn toolkit [96] version 0.19.2. The LR classifier was trained to predict, using the CKB baseline variables and findings of cardioaortic sources of embolism, a probability that each individual’s first stroke would be CE, LAA, or SAO conditional on a first IS event occurring.

F.2.3.1 Hyperparameter Tuning

Hyperparameters were tuned to maximise area under the receiver-operating characteristic curve (AUC) for each individual subtype using 5-fold cross validation in a “one-versus-rest” multiclass setup. The L1 penalty was used for regularisation and feature selection, and the LR model was computed using the “liblinear” solver.

For the main model described in the manuscript, the final tuned values for the hyperparameter, “C” (describing the inverse of regularisation strength) were: 2.78 for CE, 0.36 for LAA, and 0.36 for SAO.

For the sensitivity analysis model, which excluded silent cerebral infarction, a similar hyperparameter tuning process was implemented, and the final tuned values for “C” were: 0.36 for CE, 0.36 for LAA, and 0.36 for SAO.

F.3 Additional Figures and Tables

Table F.2: Breakdown of primary ischaemic stroke (IS) cases with separation of silent and non-silent strokes.

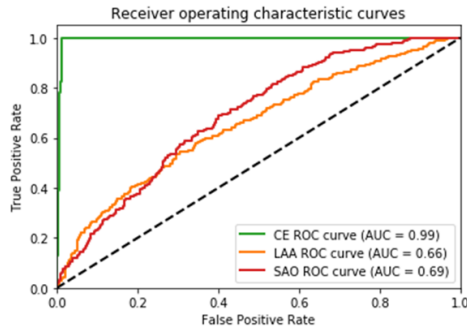
CCS Causative Subtype	Total Primary IS Cases n=22,216	Non-Silent Stroke n=18,232	Silent Cerebral Infarct n=3,984
<u>Determined Cases</u>			
CE-evident	157 (1%)	153 (1%)	4 (0%)
CE-probable	3 (0%)	3 (0%)	0 (0%)
CE-possible	235 (1%)	228 (1%)	7 (0%)
LAA-evident	888 (4%)	862 (5%)	26 (1%)
LAA-probable	1,507 (7%)	1,166 (6%)	341 (9%)
LAA-possible	1,953 (9%)	1,872 (10%)	81 (2%)
SAO-evident	4,620 (21%)	3,366 (18%)	1,254 (31%)
SAO-probable	268 (1%)	259 (1%)	9 (0%)
SAO-possible	5,900 (27%)	4,284 (23%)	1,616 (41%)
<u>Undetermined Cases</u>			
>1 identified likely cause:	1,008 (5%)	864 (5%)	144 (4%)
No identified likely cause:	5,677 (26%)	5,175 (28%)	502 (13%)

CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion.

Table F.3: Top risk factors for classifying ischaemic stroke (IS) subtypes (excluding silent cerebral infarcts).

CE		LAA		SAO	
Risk Factor	OR [95% CI]	Risk Factor	OR [95% CI]	Risk Factor	OR [95% CI]
ce_risk	131 [106-168]	Hunan	2.98 [2.21-3.94]	ce_risk	0.15 [0.11-0.19]
Children	0.68 [0.51-0.88]	Female	0.57 [0.46-0.74]	Hunan	0.32 [0.24-0.44]
		Harbin	0.58 [0.46-0.73]	Female	1.84 [1.39-2.28]
		Suzhou	1.57 [1.16-2.52]	Harbin	1.74 [1.38-2.23]
		Haikou	1.50 [1.06-1.98]	Suzhou	0.60 [0.41-0.82]
		Eats wheat monthly	0.81 [0.69-1.00]	Haikou	0.68 [0.52-0.92]
		Father stroke	0.83 [0.68-0.99]	Father stroke	1.22 [1.01-1.44]
		Household size	1.07 [1.02-1.10]	Eats wheat monthly	1.20 [1.01-1.45]
		SBP_noHRX	1.06 [1.01-1.14]	Household size	0.94 [0.90-0.98]
				Physical activity (MET-h)	1.01 [1.00-1.03]

CE: Cardioaortic embolism. LAA: Large artery atherosclerosis. SAO: Small artery occlusion. OR: Odds ratio. Risk factors ranked by magnitude of logistic regression coefficient and included if 95% CI indicated statistical significance. Red cells indicate increased likelihood of corresponding IS subtype. Green cells indicate decreased likelihood of corresponding IS subtype.

**One-vs-One ROC AUC scores:**

0.82 (macro-averaged),
0.80 (weighted by prevalence)

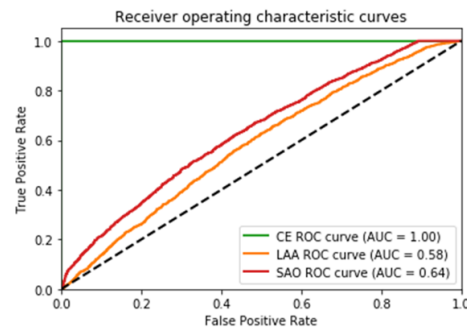
One-vs-Rest ROC AUC scores:

0.78 (macro-averaged),
0.69 (weighted by prevalence)

	ML-Classified CE	ML-Classified LAA	ML-Classified SAO
Actual CE	19	4	0
Actual LAA	11	91	203
Actual SAO	0	59	485

	PPV	Sensitivity	F1 Score	Support
CE	0.63	0.83	0.72	23
LAA	0.59	0.30	0.40	305
SAO	0.70	0.89	0.79	544
Accuracy			0.68	872
Macro-Averaged	0.64	0.67	0.63	872
Weighted Average	0.66	0.68	0.65	872

Figure F.1: Performance of multiclass logistic regression (LR) model on evident and probable cases of each ischaemic stroke (IS) subtype in test set (excluding silent cerebral infarcts).

**One-vs-One ROC AUC scores:**

0.78 (macro-averaged),
0.76 (weighted by prevalence)

One-vs-Rest ROC AUC scores:

0.74 (macro-averaged),
0.63 (weighted by prevalence)

	ML-Classified CE	ML-Classified LAA	ML-Classified SAO
Actual CE	196	31	1
Actual LAA	0	315	1557
Actual SAO	0	504	3780

	PPV	Sensitivity	F1 Score	Support
CE	1.00	0.86	0.92	228
LAA	0.37	0.17	0.23	1,872
SAO	0.71	0.88	0.79	4,284
Accuracy			0.67	6,384
Macro-Averaged	0.69	0.64	0.65	6,384
Weighted Average	0.62	0.67	0.63	6,384

Figure F.2: Performance of multiclass logistic regression (LR) model on possible cases of each ischaemic stroke (IS) subtype in test set (excluding silent cerebral infarcts).

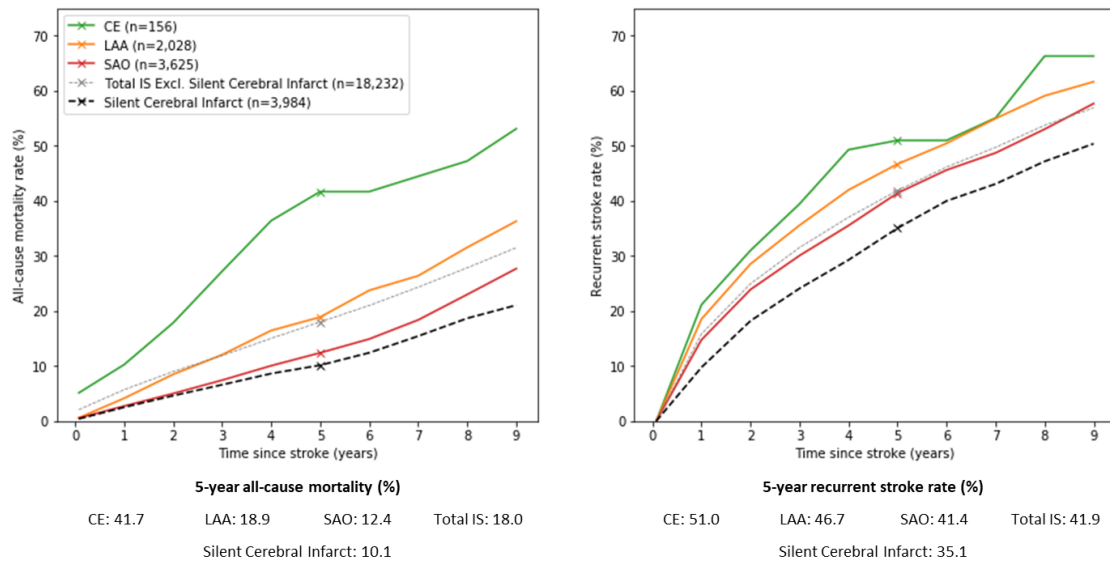


Figure F.3: Estimated cumulative all-cause mortality and recurrent stroke rates after first ischaemic stroke (IS) of (i) different subtypes (excluding silent cerebral infarcts) and (ii) silent cerebral infarcts.

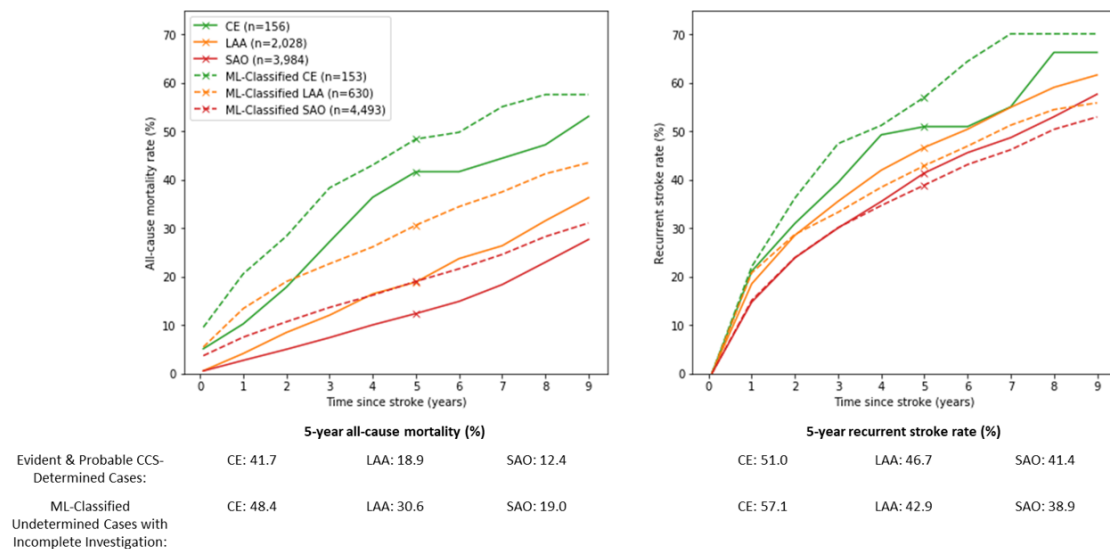


Figure F.4: All-cause mortality and recurrent stroke rates of different IS subtypes for (i) cases with evident and probable determined aetiology and (ii) ML-classified strokes with undetermined aetiology due to incomplete investigation (excluding silent cerebral infarcts).

References

1. Roth Gregory A. *et al.* Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. *Journal of the American College of Cardiology* **76**, 2982–3021 (Dec. 22, 2020).
2. Institute for Health Metrics and Evaluation (IHME). *GBD Compare* Seattle, WA: IHME, University of Washington, 2019.
3. Gorelick, P. B. The global burden of stroke: persistent and disabling. *The Lancet Neurology* **18**, 417–418 (May 1, 2019).
4. Brainin, M. *et al.* Cut stroke in half: Polypill for primary prevention in stroke. *Int J Stroke* **13**, 633–647 (Aug. 2018).
5. Kalkonde, Y. V., Alladi, S., Kaul, S. & Hachinski, V. Stroke Prevention Strategies in the Developing World. *Stroke* **49**, 3092–3097 (Dec. 1, 2018).
6. Johnson, W., Onuma, O., Owolabi, M. & Sachdev, S. Stroke: a global response is needed. *Bull World Health Organ* **94**, 634–634A (Sept. 1, 2016).
7. Wang Wenzhi *et al.* Prevalence, Incidence, and Mortality of Stroke in China. *Circulation* **135**, 759–771 (Feb. 21, 2017).
8. Dufouil Carole *et al.* Revised Framingham Stroke Risk Profile to Reflect Temporal Trends. *Circulation* **135**, 1145–1159 (Mar. 21, 2017).
9. 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* **346**, f2573 (May 2, 2013).
10. Xing Xiaolong *et al.* Predicting 10-Year and Lifetime Stroke Risk in Chinese Population. *Stroke* **50**, 2371–2378 (Sept. 1, 2019).
11. Ferket, B. S. *et al.* Separate prediction of intracerebral hemorrhage and ischemic stroke. *Neurology* **82**, 1804–1812 (May 20, 2014).
12. Hatleberg, C. I. *et al.* Predictors of Ischemic and Hemorrhagic Strokes Among People Living With HIV: The D:A:D International Prospective Multicohort Study. *EClinicalMedicine* **13**, 91–100 (Aug. 1, 2019).
13. Adams H P *et al.* Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* **24**, 35–41 (Jan. 1, 1993).
14. Ay, H. *et al.* A computerized algorithm for etiologic classification of ischemic stroke: the Causative Classification of Stroke System. *Stroke* **38**, 2979–2984 (Nov. 2007).
15. Aho, K. *et al.* Cerebrovascular disease in the community: results of a WHO collaborative study. *Bull World Health Organ* **58**, 113–130 (1980).

16. Coupland, A., Thapar, A., Qureshi, M., Jenkins, H. & Davies, A. The definition of stroke. *Journal of the Royal Society of Medicine* **110**, 9–12 (Jan. 1, 2017).
17. Sacco Ralph L. *et al.* An Updated Definition of Stroke for the 21st Century. *Stroke* **44**, 2064–2089 (July 1, 2013).
18. Brainin, M. & Heiss, W.-D. *Textbook of Stroke Medicine* 475 pp. (Cambridge University Press, Apr. 18, 2019).
19. Chen, Z. *et al.* Adiposity and risk of ischaemic and haemorrhagic stroke in 0·5 million Chinese men and women: a prospective cohort study. *The Lancet Global Health* **6**, e630–e640 (June 1, 2018).
20. Sun, L. *et al.* Causal associations of blood lipids with risk of ischaemic stroke and intracerebral haemorrhage in Chinese adults. *Nat Med* **25**, 569–574 (Apr. 1, 2019).
21. Song, Y.-M. *et al.* Blood pressure, haemorrhagic stroke, and ischaemic stroke: the Korean national prospective occupational cohort study. *BMJ* **328**, 324–325 (Feb. 5, 2004).
22. Zia Elisabet *et al.* Blood Pressure in Relation to the Incidence of Cerebral Infarction and Intracerebral Hemorrhage. *Stroke* **38**, 2681–2685 (Oct. 1, 2007).
23. Hemphill J. Claude *et al.* Guidelines for the Management of Spontaneous Intracerebral Hemorrhage. *Stroke* **46**, 2032–2060 (July 1, 2015).
24. Ay, H. *et al.* An evidence-based causative classification system for acute ischemic stroke. *Annals of Neurology* **58**, 688–697 (2005).
25. Amarenco, P. *et al.* The ASCOD Phenotyping of Ischemic Stroke (Updated ASCO Phenotyping). *CED* **36**, 1–5 (2013).
26. Amarenco, P., Bogousslavsky, J., Caplan, L. R., Donnan, G. A. & Hennerici, M. G. New approach to stroke subtyping: the A-S-C-O (phenotypic) classification of stroke. *Cerebrovasc Dis* **27**, 502–508 (2009).
27. Tsai, C.-F., Thomas, B. & Sudlow, C. L. Epidemiology of stroke and its subtypes in Chinese vs white populations. *Neurology* **81**, 264–272 (July 16, 2013).
28. World Health Organization. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision* (World Health Organization, 1992).
29. Go Alan S. *et al.* Heart Disease and Stroke Statistics—2014 Update. *Circulation* **129**, e28–e292 (Jan. 21, 2014).
30. Meschia James F. *et al.* Guidelines for the Primary Prevention of Stroke. *Stroke* **45**, 3754–3832 (Dec. 1, 2014).
31. Wang, Y. *et al.* Chinese Stroke Association guidelines for clinical management of cerebrovascular disorders: executive summary and 2019 update of the management of high-risk population. *Stroke Vasc Neurol* **5** (Sept. 1, 2020).
32. O'Donnell, M. J. *et al.* Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *The Lancet* **376**, 112–123 (July 10, 2010).
33. Parascandola, M. & Xiao, L. Tobacco and the lung cancer epidemic in China. *Transl Lung Cancer Res* **8**, S21–S30 (Suppl 1 May 2019).

34. Feigin, V. L. & Norrving, B. A new paradigm for primary prevention strategy in people with elevated risk of stroke. *Int J Stroke* **9**, 624–626 (July 2014).
35. Brainin, M. *et al.* Multi-level community interventions for primary stroke prevention: A conceptual approach by the World Stroke Organization. *International Journal of Stroke* **14**, 818–825 (Oct. 1, 2019).
36. Guan, T. *et al.* Rapid transitions in the epidemiology of stroke and its risk factors in China from 2002 to 2013. *Neurology* **89**, 53–61 (July 4, 2017).
37. Jiang, Y., Kong, L. & Li, L. Implementing the strategy of ‘Healthy China’ and strengthening the setting-up of National Demonstration Areas, for Comprehensive Prevention and Control of Non-communicable Diseases. *Zhonghua Liu Xing Bing Xue Za Zhi* **39**, 391–393 (Apr. 1, 2018).
38. Zhang, J. *et al.* Multilevel evaluation of ‘China Healthy Lifestyles for All’, a nationwide initiative to promote lower intakes of salt and edible oil. *Preventive Medicine* **67**, 210–215 (Oct. 1, 2014).
39. Li, Z., Jiang, Y., Li, H., Xian, Y. & Wang, Y. China’s response to the rising stroke burden. *BMJ* **364**, 1879 (Feb. 28, 2019).
40. Chinese Center for Disease Control and Prevention. *China National Plan for NCD Prevention and Treatment (2012-2015)* (May 8, 2012).
41. Li Zixiao *et al.* Substantial Progress Yet Significant Opportunity for Improvement in Stroke Care in China. *Stroke* **47**, 2843–2849 (Nov. 1, 2016).
42. Wang, Y. *et al.* Stroke care quality in China: Substantial improvement, and a huge challenge and opportunity. *International Journal of Stroke* **12**, 229–235 (Apr. 1, 2017).
43. Wang, Y. *et al.* Rationale and design of a cluster-randomized multifaceted intervention trial to improve stroke care quality in China: The GOLDEN BRIDGE–Acute Ischemic Stroke. *American Heart Journal* **169**, 767–774.e2 (June 1, 2015).
44. Wald, N. J. & Law, M. R. A strategy to reduce cardiovascular disease by more than 80%. *BMJ* **326**, 1419 (June 26, 2003).
45. Liu, M. *et al.* Stroke in China: epidemiology, prevention, and management strategies. *The Lancet Neurology* **6**, 456–464 (May 1, 2007).
46. Goff David C. *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* **129**, S49–S73 (June 24, 2014).
47. Hippisley-Cox, J., Coupland, C. & Brindle, P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* **357**, j2099 (May 23, 2017).
48. Yang Xueli *et al.* Predicting the 10-Year Risks of Atherosclerotic Cardiovascular Disease in Chinese Population. *Circulation* **134**, 1430–1440 (Nov. 8, 2016).
49. Anderson, K. M., Odell, P. M., Wilson, P. W. F. & Kannel, W. B. Cardiovascular disease risk profiles. *American Heart Journal* **121**, 293–298 (Jan. 1, 1991).
50. Wolf P A, D’Agostino R B, Belanger A J & Kannel W B. Probability of stroke: a risk profile from the Framingham Study. *Stroke* **22**, 312–318 (Mar. 1, 1991).

51. D'Agostino R B, Wolf P A, Belanger A J & Kannel W B. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke* **25**, 40–43 (Jan. 1, 1994).
52. Moons, K. *et al.* Prediction of stroke in the general population in Europe (EUROSTROKE): Is there a role for fibrinogen and electrocardiography? *J Epidemiol Community Health* **56**, i30–i36 (Suppl 1 Feb. 2002).
53. Chambless, L. E., Heiss, G., Shahar, E., Earp, M. J. & Toole, J. Prediction of Ischemic Stroke Risk in the Atherosclerosis Risk in Communities Study. *American Journal of Epidemiology* **160**, 259–269 (Aug. 1, 2004).
54. Jee, S. H. *et al.* Stroke risk prediction model: A risk profile from the Korean study. *Atherosclerosis* **197**, 318–325 (Mar. 1, 2008).
55. Chien Kuo-Liong *et al.* Constructing the Prediction Model for the Risk of Stroke in a Chinese Population. *Stroke* **41**, 1858–1864 (Sept. 1, 2010).
56. Zhou, X.-H., Wang, X., Duncan, A., Hu, G. & Zheng, J. Statistical evaluation of adding multiple risk factors improves Framingham stroke risk score. *BMC Med Res Methodol* **17** (Apr. 14, 2017).
57. Gage, B. F. *et al.* Validation of Clinical Classification Schemes for Predicting Stroke: Results From the National Registry of Atrial Fibrillation. *JAMA* **285**, 2864 (June 13, 2001).
58. Kothari Viti *et al.* UKPDS 60: risk of stroke in type 2 diabetes estimated by the UK Prospective Diabetes Study risk engine. *Stroke* **33**, 1776–1781 (July 1, 2002).
59. Diener, H.-C., Ringleb, P. A. & Savi, P. Clopidogrel for the secondary prevention of stroke. *Expert Opinion on Pharmacotherapy* **6**, 755–764 (May 1, 2005).
60. Johnston, S. C. *et al.* Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. *The Lancet* **369**, 283–292 (Jan. 27, 2007).
61. Lip, G. Y. H., Nieuwlaat, R., Pisters, R., Lane, D. A. & Crijns, H. J. G. M. Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor-Based Approach: The Euro Heart Survey on Atrial Fibrillation. *Chest* **137**, 263–272 (Feb. 1, 2010).
62. Khosla, A. *et al.* An integrated machine learning approach to stroke prediction in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (Association for Computing Machinery, New York, NY, USA, July 25, 2010), 183–192.
63. Letham, B., Rudin, C., McCormick, T. H. & Madigan, D. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* **9**, 1350–1371 (Sept. 2015).
64. Hung, C.-Y., Chen, W.-C., Lai, P.-T., Lin, C.-H. & Lee, C.-C. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (July 2017), 3110–3113.

65. Hung, C.-Y., Lin, C.-H., Lan, T.-H., Peng, G.-S. & Lee, C.-C. Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database. *PLOS ONE* **14**, e0213007 (Mar. 13, 2019).
66. Li, X., Bian, D., Yu, J., Li, M. & Zhao, D. Using machine learning models to improve stroke risk level classification methods of China national stroke screening. *BMC Medical Informatics and Decision Making* **19**, 261 (Dec. 10, 2019).
67. Orfanoudaki, A. *et al.* Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score. *PLOS ONE* **15**, e0232414 (May 21, 2020).
68. Truelsen T, Lindenstrøm E & Boysen G. Comparison of probability of stroke between the Copenhagen City Heart Study and the Framingham Study. *Stroke* **25**, 802–807 (Apr. 1, 1994).
69. McClure Leslie A. *et al.* Assessing the Performance of the Framingham Stroke Risk Score in the Reasons for Geographic and Racial Differences in Stroke Cohort. *Stroke* **45**, 1716–1720 (June 1, 2014).
70. Flueckiger Peter, Longstreth Will, Herrington David & Yeboah Joseph. Revised Framingham Stroke Risk Score, Nontraditional Risk Markers, and Incident Stroke in a Multiethnic Cohort. *Stroke* **49**, 363–369 (Feb. 1, 2018).
71. Leung, J. Y., Lin, S. L., Lee, R. S., Lam, T. H. & Schooling, C. M. Framingham risk score for predicting cardiovascular disease in older adults in Hong Kong. *Hong Kong Med J* **24 Suppl 4**, 8–11 (Aug. 2018).
72. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE* **12**, e0174944 (Apr. 4, 2017).
73. Yedavalli, V. S., Tong, E., Martin, D., Yeom, K. W. & Forkert, N. D. Artificial intelligence in stroke imaging: Current and future perspectives. *Clinical Imaging* **69**, 246–254 (Jan. 1, 2021).
74. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. *Doctor AI: Predicting Clinical Events via Recurrent Neural Networks in Machine Learning for Healthcare Conference* (PMLR, Dec. 10, 2016), 301–318.
75. McDermott, M. B. A. *et al.* Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine* **13** (Mar. 24, 2021).
76. Diprose, W. K. *et al.* Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association* **27**, 592–600 (Apr. 1, 2020).
77. Collins Gary S., Reitsma Johannes B., Altman Douglas G. & Moons Karel G.M. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). *Circulation* **131**, 211–219 (Jan. 13, 2015).
78. Chen, Z. *et al.* Cohort Profile: The Kadoorie Study of Chronic Disease in China (KSCDC). *International Journal of Epidemiology* **34**, 1243–1249 (Dec. 1, 2005).
79. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology* **40**, 1652–1666 (Dec. 1, 2011).

80. Chen, Y. *et al.* Mortality and recurrent vascular events after first incident stroke: a 9-year community-based study of 0·5 million Chinese adults. *Lancet Glob Health* **8**, e580–e590 (Apr. 2020).
81. Fan, M. *et al.* Family History, Tobacco Smoking, and Risk of Ischemic Stroke. *J Stroke* **21**, 175–183 (May 2019).
82. Xue, X. *et al.* Testing the proportional hazards assumption in case-cohort analysis. *BMC Medical Research Methodology* **13**, 88 (July 9, 2013).
83. Klein, J. P., Houwelingen, H. C. v., Ibrahim, J. G. & Scheike, T. H. *Handbook of Survival Analysis* 635 pp. (CRC Press, Apr. 19, 2016).
84. Breslow, N. Discussion on Professor Cox’s Paper. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 202–220 (1972).
85. Lin, D. Y. On the Breslow estimator. *Lifetime Data Anal* **13**, 471–480 (Dec. 1, 2007).
86. Friedman, J., Hastie, T. & Tibshirani, R. *Glmnet: lasso and elastic-net regularized generalized linear models*. *R package* version 3.0-2. 2019.
87. Hastie, T. & Qian, J. *Glmnet vignette* 2014.
88. Holmes, M. V. *et al.* Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *J Am Coll Cardiol* **71**, 620–632 (Feb. 13, 2018).
89. Dong, W. *et al.* Self-Rated Health Status and Risk of Incident Stroke in 0.5 Million Chinese Adults: The China Kadoorie Biobank Study. *J Stroke* **20**, 247–257 (May 2018).
90. Moons, K. G. M. *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* **98**, 683–690 (May 2012).
91. D’Agostino, R. B. & Nam, B.-H. in *Handbook of Statistics* 1–25 (Elsevier, Jan. 1, 2003).
92. Demler, O. V., Paynter, N. P. & Cook, N. R. Tests of Calibration and Goodness of Fit in the Survival Setting. *Stat Med* **34**, 1659–1680 (May 10, 2015).
93. Fine, J. P. & Gray, R. J. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* **94**, 496–509 (1999).
94. Gerds, T. *et al.* *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*. *R package* version 2020.12.08. 2020.
95. Davidson-Pilon, C. *et al.* *CamDavidsonPilon/lifelines* version 0.21.1.
96. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
97. Damen, J. A. *et al.* Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis. *BMC Medicine* **17**, 109 (June 13, 2019).
98. Pennells, L. *et al.* Equalization of four cardiovascular risk algorithms after systematic recalibration: individual-participant meta-analysis of 86 prospective studies. *Eur Heart J* **40**, 621–631 (Feb. 14, 2019).

99. Lacey, B. *et al.* Age-specific association between blood pressure and vascular and non-vascular chronic diseases in 0·5 million adults in China: a prospective cohort study. *Lancet Glob Health* **6**, e641–e649 (June 2018).
100. Lewington, S. *et al.* Seasonal variation in blood pressure and its relationship with outdoor temperature in 10 diverse regions of China: the China Kadoorie Biobank. *J Hypertens* **30**, 1383–1391 (July 2012).
101. Arnett Donna K. *et al.* 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **140**, e596–e646 (Sept. 10, 2019).
102. Joint Committee for Guideline Revision. 2016 Chinese guidelines for the management of dyslipidemia in adults. *J Geriatr Cardiol* **15**, 1–29 (Jan. 2018).
103. Joint Committee for Guideline Revision. 2018 Chinese Guidelines for Prevention and Treatment of Hypertension-A report of the Revision Committee of Chinese Guidelines for Prevention and Treatment of Hypertension. *J Geriatr Cardiol* **16**, 182–241 (Mar. 2019).
104. Zhou, Z. & Hu, D. An epidemiological study on the prevalence of atrial fibrillation in the Chinese population of mainland China. *J Epidemiol* **18**, 209–216 (2008).
105. Guo, Y. *et al.* Prevalence, incidence, and lifetime risk of atrial fibrillation in China: new insights into the global burden of atrial fibrillation. *Chest* **147**, 109–119 (Jan. 2015).
106. Wang, Z. *et al.* The Disease Burden of Atrial Fibrillation in China from a National Cross-sectional Survey. *Am J Cardiol* **122**, 793–798 (Sept. 1, 2018).
107. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* 1102 pp. (MIT Press, Sept. 7, 2012).
108. Niculescu-Mizil, A. & Caruana, R. *Predicting good probabilities with supervised learning in Proceedings of the 22nd international conference on Machine learning* (Association for Computing Machinery, New York, NY, USA, Aug. 7, 2005), 625–632.
109. Breiman, L. *Classification and Regression Trees* 369 pp. (Routledge, Oct. 19, 2017).
110. Friedman, J. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis* **38**, 367–378 (Feb. 1, 2002).
111. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* 801 pp. (MIT Press, Nov. 18, 2016).
112. Gori, M. & Tesi, A. On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**, 76–86 (Jan. 1992).
113. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics* **2**, 841–860 (Sept. 2008).
114. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, 1–17 (Mar. 31, 2017).
115. Chollet, F. *Keras* version 2.3.1. 2015.

116. Krijthe, J. *Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation*. R package version 0.15. 2015.
117. Sun, L. *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLOS Medicine* **18**, e1003498 (Jan. 14, 2021).
118. Stone, N. J. *et al.* 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults. *Circulation* **129**, S1–S45 (June 24, 2014).
119. National Institute for Health and Care Excellence. *CVD risk assessment and management: Scenario: Management of people with an estimated risk of 10% or more* Mar. 2019.
120. De Vries, T. I. & Visseren, F. L. J. Cardiovascular risk prediction tools made relevant for GPs and patients. *Heart* **107**, 332–340 (Feb. 1, 2021).
121. Graham, I. *et al.* European guidelines on cardiovascular disease prevention in clinical practice: executive summary: Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by representatives of nine societies and by invited experts). *Eur Heart J* **28**, 2375–2414 (Oct. 2007).
122. Kavousi, M. *et al.* Comparison of application of the ACC/AHA guidelines, Adult Treatment Panel III guidelines, and European Society of Cardiology guidelines for cardiovascular disease prevention in a European cohort. *JAMA* **311**, 1416–1423 (Apr. 9, 2014).
123. Nuotio, J., Suvila, K., Cheng, S., Langén, V. & Niiranen, T. Longitudinal blood pressure patterns and cardiovascular disease risk. *Ann Med* **52**, 43–54 (June 2020).
124. Sweeting, M. J., Barrett, J. K., Thompson, S. G. & Wood, A. M. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the ARIC study. *Stat Med* **36**, 4514–4528 (Dec. 10, 2017).
125. Paynter, N. P., Crainiceanu, C. M., Sharrett, A. R., Chambless, L. E. & Coresh, J. Effect of correcting for long-term variation in major coronary heart disease risk factors: relative hazard estimation and risk prediction in the Atherosclerosis Risk in Communities Study. *Ann Epidemiol* **22**, 191–197 (Mar. 2012).
126. Paige, E. *et al.* Use of Repeated Blood Pressure and Cholesterol Measurements to Improve Cardiovascular Disease Risk Prediction: An Individual-Participant-Data Meta-Analysis. *American Journal of Epidemiology* **186**, 899–907 (Oct. 15, 2017).
127. Cho, I.-J., Sung, J. M., Chang, H.-J., Chung, N. & Kim, H. C. Incremental Value of Repeated Risk Factor Measurements for Cardiovascular Disease Prediction in Middle-Aged Korean Adults: Results From the NHIS-HEALS (National Health Insurance System-National Health Screening Cohort). *Circ Cardiovasc Qual Outcomes* **10** (Nov. 2017).
128. Zhao, J. *et al.* Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Sci Rep* **9**, 717 (Jan. 24, 2019).
129. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1–48 (Oct. 7, 2015).

130. Chun, M. *et al.* Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *Journal of the American Medical Informatics Association* (ocab068 May 9, 2021).
131. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (Sept. 1988).
132. Peto, R. The Horse-Racing Effect. *The Lancet* **318**, 467–468 (Aug. 29, 1981).
133. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* **15**, 361–387 (Feb. 28, 1996).
134. Burch, P. M. *et al.* Net Reclassification Index and Integrated Discrimination Index Are Not Appropriate for Testing Whether a Biomarker Improves Predictive Performance. *Toxicol Sci* **156**, 11–13 (Mar. 2017).
135. Yang, W. *et al.* Serum lipids and lipoproteins in Chinese men and women. *Circulation* **125**, 2212–2221 (May 8, 2012).
136. Cholesterol Treatment Trialists' (CTT) Collaboration. Efficacy and safety of LDL-lowering therapy among men and women: meta-analysis of individual data from 174,000 participants in 27 randomised trials. *The Lancet* **385**, 1397–1405 (Apr. 11, 2015).
137. Catapano, A. L. *et al.* 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias. *European Heart Journal* **37**, 2999–3058 (Oct. 14, 2016).
138. Liu, Y. *et al.* Time trends analysis of statin prescription prevalence, therapy initiation, dose intensity, and utilization from the hospital information system of Jinshan Hospital, Shanghai (2012–2018). *BMC Cardiovasc Disord* **20**, 201 (Apr. 25, 2020).
139. Yusuf, S. *et al.* Use of secondary prevention drugs for cardiovascular disease in the community in high-income, middle-income, and low-income countries (the PURE Study): a prospective epidemiological survey. *The Lancet* **378**, 1231–1243 (Oct. 1, 2011).
140. Wang, M. *et al.* Impact of China's Low Centralized Medicine Procurement Prices on the Cost-Effectiveness of Statins for the Primary Prevention of Atherosclerotic Cardiovascular Disease. *Glob Heart* **15**, 43 (June 25, 2020).
141. Stevens, W. *et al.* Estimating the future burden of cardiovascular disease and the value of lipid and blood pressure control therapies in China. *BMC Health Serv Res* **16**, 175 (May 10, 2016).
142. Tomlinson, B., Chan, P. & Liu, Z.-M. Statin Responses in Chinese Patients. *J Atheroscler Thromb* **25**, 199–202 (Feb. 1, 2018).
143. Zhou, M. *et al.* Cause-specific mortality for 240 causes in China during 1990–2013: a systematic subnational analysis for the Global Burden of Disease Study 2013. *The Lancet* **387**, 251–272 (Jan. 16, 2016).
144. Zhao, D. *et al.* Epidemiological transition of stroke in China: twenty-one-year observational study from the Sino-MONICA-Beijing Project. *Stroke* **39**, 1668–1674 (June 2008).

145. Collins, R. *et al.* Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet* **388**, 2532–2561 (Nov. 19, 2016).
146. Zhang, X.-F., Attia, J., D’Este, C., Yu, X.-H. & Wu, X.-G. A risk score predicted coronary heart disease and stroke in a Chinese cohort. *J Clin Epidemiol* **58**, 951–958 (Sept. 2005).
147. Lip, G. Y. H. *et al.* Comparative assessment of the HAS-BLED score with other published bleeding risk scoring schemes, for intracranial haemorrhage risk in a non-atrial fibrillation population: the Chin-Shan Community Cohort Study. *Int J Cardiol* **168**, 1832–1836 (Oct. 3, 2013).
148. Tsai, C.-F., Anderson, N., Thomas, B. & Sudlow, C. L. M. Risk factors for ischemic stroke and its subtypes in Chinese vs. Caucasians: Systematic review and meta-analysis. *Int J Stroke* **10**, 485–493 (June 2015).
149. Tan, Y.-F. *et al.* Risk Factors, Clinical Features and Prognosis for Subtypes of Ischemic Stroke in a Chinese Population. *Curr Med Sci* **38**, 296–303 (Apr. 2018).
150. Tian, D. *et al.* Trends in stroke subtypes and vascular risk factors in a stroke center in China over 10 years. *Sci Rep* **8**, 5037 (Mar. 22, 2018).
151. Arsava, E. M. *et al.* Assessment of the Predictive Validity of Etiologic Stroke Classification. *JAMA Neurol* **74**, 419–426 (Apr. 1, 2017).
152. World Health Organization. *ICD-11 : international statistical classification of diseases and related health problems : eleventh revision* (World Health Organization, 2020).
153. Li, L. *et al.* Incidence, outcome, risk factors, and long-term prognosis of cryptogenic transient ischaemic attack and ischaemic stroke: a population-based study. *Lancet Neurol* **14**, 903–913 (Sept. 2015).
154. Ay, H. *et al.* Pathogenic ischemic stroke phenotypes in the NINDS-stroke genetics network. *Stroke* **45**, 3589–3596 (Dec. 2014).
155. Garg, R., Oh, E., Naidech, A., Kording, K. & Prabhakaran, S. Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. *J Stroke Cerebrovasc Dis* **28**, 2045–2051 (July 2019).
156. Fang, G., Xu, P. & Liu, W. Automated Ischemic Stroke Subtyping Based on Machine Learning Approach. *IEEE Access* **8**, 118426–118432 (2020).
157. Guan, W. *et al.* Automated Electronic Phenotyping of Cardioembolic Stroke. *Stroke* **52**, 181–189 (Jan. 2021).
158. Cyranoski, D. China embraces precision medicine on a massive scale. *Nature* **529**, 9–10 (Jan. 1, 2016).
159. Hadley, T. D., Agha, A. M. & Ballantyne, C. M. How Do We Incorporate Polygenic Risk Scores in Cardiovascular Disease Risk Assessment and Management? *Curr Atheroscler Rep* **23**, 28 (Apr. 1, 2021).