University of Nottingham
UK | CHINA | MALAYSIA

Stratifying Stroke Severity: Towards a Personalised Medicine Application for Primary Care

Ralph Kwame Akyea
BSc (Med Sci), MBChB, MPH

Centre for Academic Primary Care
Lifespan and Population Health Unit
School of Medicine

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

January 2022

To:

Yoofi, Ebow, and Maame Araba

Abstract

Stroke remains a major cause of death and disability worldwide, despite advances in prevention and treatment. Improvements in acute care have led to many surviving after an incident stroke event. However, the prognosis after surviving remains compromised. This is due to the high risk of recurrent adverse cardiovascular events, greatest during the first year but persisting over one's lifetime. Reducing long-term residual cardiovascular risk and improving quality of life are primary goals for clinical practice and research. Identifying patients at the greatest risk of subsequent major adverse cardiovascular events (MACE) could help clinicians and policymakers determine which patients need to be prioritised.

This thesis research aimed to identify clinical phenotypes (that is, patient characteristics and distinct patient clusters) that correlate with subsequent MACE outcomes (defined as a diagnosis of either CHD, recurrent stroke, PVD, heart failure, or CVD-related mortality) in adults with an incident stroke diagnosis.

Firstly, a systematic review was completed to identify and summarise the available evidence on prognostic models and assess their accuracy for predicting MACE outcomes in an adult with established stroke. Forty (40) full-text articles with 23 distinct prognostic models for predicting MACE outcomes in adults with established stroke were identified by the systematic review. There were 11 prognostic model developments and 77 external validations of models reported. Among the 23 models, the most frequently used predictors were age, sex, history of transient ischaemic attack, hypertension (blood pressure), and diabetes. Critical appraisal

identified methodological limitations, in particular: inadequate sample size, improper handling of missing data, and incomplete evaluation of model performance.

The Clinical Practice Research Datalink (CPRD GOLD), a longitudinal database of anonymised electronic health records (UK primary care data) linked to Hospital Episode Statistics (HES APC), national death registry, and social deprivation data was then used to undertake a series of data-related studies. Four cohort studies were completed using patients aged ≥18 years with an incident stroke diagnosis between 1 January 1998 and 31 December 2017, and no prior history of either CHD, PVD or heart failure, to assess the risk of subsequent cardiovascular morbidity and mortality outcomes.

In the analysis of 9,997,376 individual records in CPRD GOLD database, there were 82,774 non-fatal incident stroke events recorded in either primary care or hospital data – a stroke incident rate of 109.20 per 100,000 person-years (95% CI: 108.46 – 109.95). Of the 82,774 patients, 13,879 (16.8%) patients had a prior history of major adverse outcomes (CHD, PVD, and heart failure) and were excluded. Subsequent MACE was recorded in 47,500 (69.0%) of the remaining 68,877 patients. In the UK, the incidence of stroke and subsequent major adverse cardiovascular morbidity and mortality outcomes were higher in women, older populations, and people living in socially deprived areas.

After excluding patients with stroke not-otherwise specified (n=36,551) and adjusting for potential confounders, patients with incident haemorrhagic stroke (n=6,535, 20.4%) had no significantly different risk of subsequent cardiovascular morbidity, compared with patients with incident ischaemic stroke (n=25,556, 79.6%) – CHD [HR 0.86, 95% CI 0.56 – 1.32], recurrent stroke [HR 0.92, 95% CI 0.83 – 1.02], PVD [HR 1.15, 95% CI 0.56 – 2.38], or heart failure [HR 1.03, 95% CI 0.61 – 1.74]. However, patients with incident haemorrhagic stroke had a

significantly higher risk of subsequent CVD-related mortality [HR 2.35, 95% CI 2.04 – 2.72] and all-cause mortality [HR 2.16, 95% CI 1.94 – 2.41]. Propensity-score matched analysis of 1,039 patients with haemorrhagic stroke and 1,039 with ischaemic stroke showed similar risk in subsequent cardiovascular morbidity outcomes – CHD, recurrent stroke, PVD and heart failure.

Obesity, a risk factor for stroke and is also a risk factor for hypertension and diabetes (known risk factors for CVD), is commonly measured using body mass index (BMI). In a multivariable analysis of a cohort of 30,702 patients with incident stroke and BMI record, individuals in higher BMI categories were associated with lower risk of subsequent:

- MACE [overweight (BMI: 25.0-29.9 kg/m^2): HR 0.96, 95% CI 0.93 – 0.99)],
- PVD [overweight: HR 0.65, 95% CI 0.49 – 0.85; obesity class III (BMI: ≥40 kg/m^2): HR 0.19, 95% CI 0.50 – 0.77],
- CVD-related mortality [overweight: HR 0.80, 95% CI 0.74 – 0.86; obesity class I (BMI: 30.0-34.9 kg/m^2): HR 0.79, 95% 0.71 – 0.88; class II (BMI: 35.0-39.9 kg/m^2): HR 0.80, 95% CI 0.67 – 0.96]; and
- all-cause mortality [overweight: HR 0.75, 95% CI 0.71 – 0.79; obesity class I: HR 0.75, 95% CI 0.70 – 0.81; class II: HR 0.77, 95% CI 0.68 – 0.86]

when compared with those with normal BMI. The results were similar irrespective of sex, smoking status, history of diabetes mellitus or cancer at the time of incident stroke.

Using a combination of data-driven feature selection approaches and clinical expert opinion, 39 out of 336 characteristics (clinical features including sociodemographic, biochemical, comorbid conditions, and prescribed medications related to stroke or CVD) at the time of incident stroke were selected. An unsupervised machine learning approach [clustering algorithm for mixed (both categorical and continuous) data] was used to identify 4 phenotypic clusters for a cohort of 48,114 patients with incident stroke and subsequent outcomes occurring

30 days after incident stroke. Cluster 1 (n=5,201, 10.8%) was a cohort with high prevalence of CHD-related risk factors and prescribed medications; cluster 2 (n=18,655, 38.8%) a cohort with low prevalence of multiple long-term conditions (MLTC); cluster 3 (n=10,244, 21.3%) a cohort with high prevalence of MLTC; and cluster 4 (n=14,014, 29.1%), the oldest population cohort and predominantly female. The phenotypic clusters had different incidences and risks for subsequent cardiovascular morbidity and mortality outcomes. For instance, the incidence of the composite outcome of recurrent stroke and CVD-related mortality was lowest in cluster 1 and highest in cluster 4 (15.13 and 23.17 per 100 person-years, respectively). The risk of subsequent recurrent stroke + CVD-related mortality was significantly increased in cluster 2 (HR 1.07, 95% CI 1.02 – 1.12); cluster 3 (HR 1.20, 95% CI 1.14 – 1.26), and cluster 4 (HR 1.29, 95% CI 1.26 – 1.33), when compared with cluster 1.

Findings from this thesis research indicate patients with incident stroke experience considerable heterogeneity in subsequent clinical outcomes. In particular, women, older patients, and those living in socially deprived areas are at greater risk of subsequent major adverse outcomes. Additionally, age at incident stroke, blood pressure, LDL cholesterol level, a diagnosis of hypertension and potency of prescribed statin were identified as key indicators of patients' phenotypic clusters and associated risk for subsequent clinical outcomes. The studies add to growing and wider evidence to identify those who may most benefit from, and be least likely to be harmed by, preventive treatment. Stratifying patients with stroke early, could lower the burden of subsequent adverse clinical outcomes, improve patients' long-term outcomes, and reduce the associated economic burden. This should, therefore, be a continuing research and public health priority.

Acknowledgements

First and foremost, I am most grateful to God for the opportunity and exceeding grace.

The work presented in this doctoral thesis was done under the supervision of Professor Nadeem Qureshi (Principal Supervisor) and Professor Joe Kai, at the Centre for Academic Primary Care, School of Medicine, University of Nottingham; Professor Folkert W. Asselbergs with Utrecht University and the University College of London; and Dr Stephen F. Weng formerly with the University of Nottingham. I am most grateful for their support and guidance throughout the entire period.

I take this opportunity to express my sincerest gratitude to everyone who supported me through my doctoral study especially all the remarkable individuals I have collaborated with.

Particular appreciation to my family – they have been phenomenal.

Funding statement

This PhD studentship was funded through the National Institute of Health Research School for Primary Care Research (NIHR SPCR) PhD Studentship and the University of Nottingham Vice Chancellor's Scholarship for Research Excellence (International) Award. The views expressed are those of the author and not necessarily those of the NIHR, the UK National Health Service, or the UK Department of Health and Social Care.

Table of Contents

List of Figures

List of Tables

Abbreviations

ACE	Angiotensin-converting-enzyme
AHA/ACC	American Heart Association and the American College of Cardiology
ALP	Alanine phosphatase
ALT	Alanine aminotransferase
AUC	Area under the curve
BMI	Body mass index
CABG	Coronary artery bypass graft
CHD	Coronary heart disease
COPD	Chronic obstructive pulmonary disease
CPRD	Clinical Practice Research Datalink
CI	Confidence interval
CT	Computer tomography
CTP	Computer tomography perfusion
CVD	Cardiovascular disease
DASH	Dietary Approaches to Stop Hypertension
DBP	Diastolic blood pressure
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DENCLUE	DENsity CLUstEring
ECG	Electrocardiogram
ESR	Erythrocyte sedimentation rate
GFR	Glomerular filtration rate
GGT	Gamma-glutamyl transferase
GP	General practitioners
HbA1c	Glycated haemoglobin
HDL	High-density lipoprotein
HES APC	Hospital Episodes Statistics Admitted Patient Care

HR	Hazard ratio
ICD	International Statistical Classification of Diseases and Health-related Problems
ICH	Intracerebral haemorrhage
IMD	Index of multiple deprivation
IQR	Interquartile range
IR	Incidence rate
IRR	Incidence rate ratio
LDL-C	Low-density lipoprotein cholesterol
LSOA	Lower super output area
MACE	Major adverse cardiovascular event
MAR	Missing at random
MCAR	Missing completely at random
MDT	Multidisciplinary team
MI	Myocardial infarction
ML	Machine learning
MNAR	Missing not at random
MRI	Magnetic resonance imaging
NCD	Non-communicable disease
NHS	National Health Service
NOS	Not otherwise specified
NSAIDS	Non-steroidal anti-inflammatory drugs
ONS	Office of National Statistics
OPCS	Office of Population Censuses and Surveys Classification of Interventions and Procedures
PCA	Principal component analysis
PCI	Percutaneous coronary intervention
PCSK9	Proprotein convertase subtilisin/kexin type 9
PRISMA	Preferred Reporting Items for Systematic Reviews & Meta-analysis
PROBAST	Prediction model Risk Of Bias ASsessment Tool
PS	Propensity score
PTCA	Percutaneous transluminal coronary angioplasty
PVD	Peripheral vascular disease
QOF	Quality Outcomes Framework
RCT	Randomised controlled trial

ROSIER	Recognition of stroke in the emergency room
rt-PA	Recombinant tissue plasminogen activator
SAH	Subarachnoid haemorrhage
SBP	Systolic blood pressure
SD	Standard deviation
SES	Socioeconomic status
sHR	Stratified hazard ratio
TG	Triglyceride
TIA	Transient ischaemic attack
TOAST	Trials of Org 10172 in Acute Stroke Treatment
TSH	Thyroid-stimulating hormone
UK	United Kingdom
WHO	World Health Organization

List of PhD-related Publications

I was involved in the conception of the various studies. I carried out all the data analyses and drafted the manuscripts for the following PhD-related peer-reviewed publications:

1. **Chapter 1**

   Akyea, R.K., Kai, J., Qureshi, N., Hamid, H. A., & Weng, S. F. (2019). Secondary Prevention of Cardiovascular Disease: Time to Rethink Stratification of Disease Severity? *European Journal of Preventive Cardiology*, 26(16), 1778-1780. https://doi.org/10.1177/2047487319850957

2. **Chapter 3**

   Akyea RK, Leonardi-Bee J, Asselbergs FW, Patel RS, Durrington P, Wierzbicki AS, Ibiwoye OH, Kai J; Qureshi N, Weng SF. (2020). Predicting major adverse cardiovascular events for secondary prevention: protocol for a systematic review and meta-analysis of risk prediction models. *BMJ Open*, 10(7), https://doi.org/10.1136/bmjopen-2019-034564

3. **Chapter 4**

   Akyea, R.K., Vinogradova, Y., Qureshi, N., Patel, R. S., Kontopantelis, E., Ntaios, G., Asselbergs F.W., Kai J., Weng, S. F. (2021). Sex, Age, and Socioeconomic Differences in Nonfatal Stroke Incidence and Subsequent Major Adverse Outcomes. *Stroke*, 52(2), 396-405. https://doi.org/10.1161/strokeaha.120.031659x

4. **Chapter 6**

   Akyea, R.K., Döhner, W., Iyen, B., Weng, S., Qureshi, N. and Ntaios, G., (2021). Obesity and long-term outcomes after incident stroke: a prospective population-based cohort study. *Journal of Cachexia, Sarcopenia and Muscle*. https://doi.org/10.1002/jcsm.12818

Manuscripts under review

- **Chapter 5**

  Akyea, R. K., Georgiopoulos, G., Iyen, B., Kai, J., Qureshi, N., Ntaios, G. Comparison of risk of serious cardiovascular events after haemorrhagic versus ischaemic stroke: a population-based study.
  Submitted to: Thrombosis and Haemostasis.

- **Chapter 7**

  Akyea, R.K., Ntaios, G., Kontopantelis, E., Georgiopoulos, G., Soria, D., Asselbergs, F.W., Kai, J., Weng, S.F., Qureshi, N. A population-based study exploring phenotypic clusters and clinical outcomes in stroke using unsupervised machine learning approach.
  Submitted to: BMC Medical Informatics and Decision Making.

List of Collaborators

The following external collaborators provided expert guidance and insight as co-authors for the peer-reviewed publications related to this PhD research:

- **Professor Wolfram Döhner**
  Professor of Stroke Research, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, BIH Center for Regenerative Therapies (BCRT) and Center for Stroke Research Berlin (CSB) Charité Universitätsmedizin Berlin, Germany

  Contribution: Interpretation of findings and critical review of study in Chapter 6. Co-author for the published work.

- **Professor Paul Durrington**
  Professor of Medicine, Cardiovascular Research Group, The University of Manchester, Manchester, UK.

  Contribution: Conceptualisation of study in Chapter 3. Co-author for the published systematic review protocol.

- **Dr Georgios Georgiopoulos**
  European Association of Cardiovascular Imaging (EACVI) Research Fellow, School of Biomedical Engineering and Imaging Sciences, St Thomas Hospital, King's College, London, UK.

  Contribution: Conceptualisation, design, interpretation of findings and critical review of study in Chapter 5. Interpretation of findings and critical review of studies in Chapter 7. Co-author for peer-reviewed papers related to these chapters.

- **Professor Evangelos Kontopantelis**
  Professor of Data Science and Health Services Research, The University of Manchester, Manchester, UK.

  Contribution: Interpretation of findings and critical review of studies in Chapters 4 & 7. Co-author for peer-reviewed papers related to these chapters.

- **Professor Jo Leonardi-Bee**
  Professor of Medicine Statistics and Epidemiology & Co-Director of the Centre for Evidence-Based Health Care, University of Nottingham, UK.

  Contribution: Conceptualisation and design of study in Chapter 3. Co-author for the published systematic review protocol.

- **Dr George Ntaios**
  Associate Professor, Department of Internal Medicine, Faculty of Medicine, School of Health Sciences, University of Thessaly, Larissa, Greece.

  Contribution: Conceptualisation, design, interpretation of findings and critical review of studies in Chapter 5 & 6. Interpretation of findings and critical review of studies in Chapters 4 & 7. Co-author for peer-reviewed papers related to these chapters.

- **Professor Riyaz S. Patel**
  Professor of Cardiology, Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, UK.

  Contribution: Conceptualisation of study in Chapter 3. Interpretation of findings and critical review of study in Chapters. Co-author for peer-reviewed papers related to these chapters.

- **Professor Anthony Wierzbicki**
  Consultant & Honorary Professor in Metabolic Medicine/Chemical Pathology, Guy's and St Thomas' NHS Foundation Trust, London, UK.

  Contribution: Conceptualisation of study in Chapter 3. Co-author for the published systematic review protocol.

Chapter 1

Introduction

This chapter provides the rationale for this thesis research, an overview of what is already known about stroke, the aim and objectives of the thesis, and finally an outline of successive chapters.

Some of the information provided in this chapter has been published in the *European Journal of Preventive Cardiology*:

Akyea, R. K., Kai, J., Qureshi, N., Hamid, H. A., & Weng, S. F. (2019). Secondary Prevention of Cardiovascular Disease: Time to Rethink Stratification of Disease Severity? European Journal of Preventive Cardiology, 26(16), 1778-1780. https://doi.org/10.1177/2047487319850957

1.1 Introduction

Stroke is a leading cause of disability and mortality, associated with increased economic burden due to acute treatment and post-stroke care.[1] Cardiovascular disease (CVD), including strokes, are less fatal as a result of various advances in patient management. As a result, a large proportion of people are living with this long-term condition, globally – with about 1.3 million stroke survivors in the United Kingdom.[2] The number of people living with stroke in Europe was estimated to increase by 27%, from 9.53 million stroke survivors in 2017 to 12.11 million in 2047.[3]

The NHS Long Term Plan[4] identifies CVD, including stroke, as a clinical priority and the single biggest condition where lives can be saved by the NHS over the next 10 years with an ambitious plan to help prevent over 150,000 CVD-related events over the period. The NHS Long Term Plan among others aims to improve and increase early detection and treatment of CVD to help patients live longer healthier lives. For patients with an established stroke event, the priority is to prevent a subsequent CVD event or death and also to improve their quality of life. A crucial step for secondary prevention is the identification of risk factors for subsequent adverse cardiovascular morbidity and mortality outcomes, their interactions, and how the variations in these factors may relate to more severe outcomes in the future. Early detection of these factors can ensure patients receive the appropriate non-pharmacological or pharmacological interventions before the disease condition worsens or complications develop, thus improving both quality of life and life expectancy.[5]

As an integral component of routine clinical care, patients' electronic health records (EHRs) contain large amounts of clinical data that could be useful for driving further secondary prevention research using innovative methodologies. EHRs provide low-cost means of accessing potentially rich longitudinal data on

large populations at the granular level of patients, across different types of health care settings.

There is also increasing potential for research to improve the accuracy of risk stratification for secondary prevention. By interrogating the large volumes of clinical data, patients with CVD can be stratified or clustered into different risk groups according to risk factors or clinical characteristics. Rather than predict the overall risk of stroke or composite measures [such as major adverse cardiovascular events (MACE)], researchers are now able to predict specific subtypes of disease,[6] such as CVD or even focused disease areas, for example, subtypes of stroke. For instance, Schiele et al,[7] emphasized the importance of carefully selecting patients with stable coronary heart disease (CHD), with high residual risks but low therapeutic risks, for intensive secondary prevention therapy. A combination of antithrombotic and lipid-lowering medications is effective in reducing ischemic events and CVD mortality but may carry additional haemorrhagic risk. Using advanced data science approaches to interrogate EHRs offer the opportunity to profile the unique characteristics of this specific group of patients, ensuring appropriate stratification, and therapy is made available to those with the greatest need and greatest benefit.

1.2 Background

Health and well-being are key priorities for most people; however, these are constantly challenged by diseases and illnesses. Even though some of these diseases are fatal, others can be prevented or treated, or their effects can be minimised if detected early. Diseases that challenge health and wellbeing can be categorised as communicable or non-communicable. Communicable diseases refer to a group of illnesses that are spread by infectious agents such as viruses or bacteria to one another.[8] Non-communicable – or chronic – diseases (NCDs),

however, are diseases that are not transmissible directly from one person to another. The four main types of NCDs are cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes. NCDs, are the leading cause of death globally, representing 63% of all annual deaths.[9] The highest proportion of NCD-related deaths (44%) are due to cardiovascular disease (CVD). This makes CVD one of the priority areas for research on prevention. Cardiovascular disease, however, is a general term that refers to the range of disorders/conditions that affect the heart and arterial blood vessels. These include:

- coronary heart disease – a disease of blood vessels supplying the heart muscles.
- cerebrovascular disease (that is, transient ischaemic attack (TIA) and strokes) – a disease of the blood vessels supplying the brain.
- peripheral vascular disease (PVD) – a disease of blood vessels supplying the arms and legs.
- rheumatic heart disease – damage to the heart muscle and heart valves from rheumatic fever caused by streptococcal bacterial infection.
- congenital heart disease – birth defects of the heart structure that affect the normal development and function of the heart.

This doctoral project focuses on CVD (specifically, patients with stroke).

1.2.1 Definition of stroke

Cerebrovascular disease refers to any one of several disorders which compromises blood supply to the underlying tissues of the brain leading to varying degrees of neurological deficits.[10]

The definition of stroke has evolved throughout the course of medicine. Initially referred to as 'apoplexy', meaning to 'strike down violently'. It referenced a sudden fall and loss of consciousness but with retained vital signs.[10] The term 'stroke' was later explained to mean an acute episode in the brain vasculature

presumed to cause clinical symptoms of neural dysfunction.[11] The World Health Organization (WHO) defines stroke as an accident to the brain with:

"*rapidly developing clinical signs of focal or global disturbance to cerebral function, with symptoms lasting 24 hours or longer, or leading to death, with no apparent cause other than of vascular origin and includes cerebral infarct, intracerebral haemorrhage, and subarachnoid haemorrhage.*"[12]

This definition excludes neurological conditions that last less than 24 hours such as transient ischaemic attack (TIA) or have apparent causes other than a vascular origin.

There are, however, two main types of strokes[13]:

- Cerebral infarct (commonly referred to as ischaemic stroke), the commonest type of stroke, accounts for about 85% of all acute strokes.[14] Ischaemic stroke occurs when the blood flow through the artery to the brain is blocked. Trial of Org 10172 in Acute Stroke Treatment (TOAST) classification is the most widely used causative classification system for acute ischaemic stroke.[15] According to TOAST there are 5 main subtypes of ischaemic strokes[16]: large vessel atherosclerosis, small vessel diseases (lacunar infarcts), cardioembolic, stroke of other determined aetiology, and stroke of undetermined aetiology (cryptogenic stroke).
- Haemorrhagic stroke, an acute bleeding from a blood vessel within the brain, accounts for about 15% of all acute strokes. There are 2 main haemorrhagic stroke subtypes: intracerebral haemorrhage and subarachnoid haemorrhage.

1.2.2 Diagnosis of stroke

A diagnosis of stroke is made based on adequate history and physical examination, complemented with diagnostic tests and neuroimaging. It begins with the identification of some focal or global neurological deficit.[17] Clinical diagnoses have

been improved using blood biomarkers in hyperacute settings, though clearly differentiating strokes and stroke mimics has proven to be a challenge[18]. Common clinical symptoms of stroke include a sudden change in speech, visual loss, diplopia (double vision), numbness or tingling, paralysis or weakness and non-orthostatic dizziness'.[19] In practice, a diagnosis of stroke hinges on interpreting a myriad of clinical signs and symptoms and does not depend on any one specific sign or symptom.[20] Due to the specifics of stroke therapy, it is necessary to obtain an accurate clinical diagnosis of the condition. Early identification and management reduce mortality and improve long-term prognosis.[21]

Diagnostic investigation or assessment for stroke include blood tests (such as serum electrolytes, renal function test, complete blood count, and other haematological tests), electrocardiograph, and non-contrast computed tomography (CT) scans. Advanced imaging methods such as computed tomography perfusion (CTP) and magnetic resonance imaging (MRI) techniques have seen increasing use in identifying patients with acute ischaemic stroke who are eligible for treatment with endovascular thrombectomy or intravenous thrombolysis.[17,22] Diagnosis of acute stroke can also be augmented with risk prediction algorithms or models, assisting in the elimination of stroke mimics and identifying key comorbidities prior to treatment/management[23] – Figure 1.1.

There is a broad range of differential diagnoses which includes stroke mimics such as TIA[1], metabolic derangement (hypoglycaemia, hyponatremia), hemiplegic migraine, abscesses from infections, brain tumour, syncope, and conversion disorder.[24]

[1] TIA, commonly referred to as a mini stroke, is defined as a transient episode of neurologic dysfunction caused by focal cerebral, spinal cord, or retinal ischemia, without acute infarction (Simmons, et al., 2012).

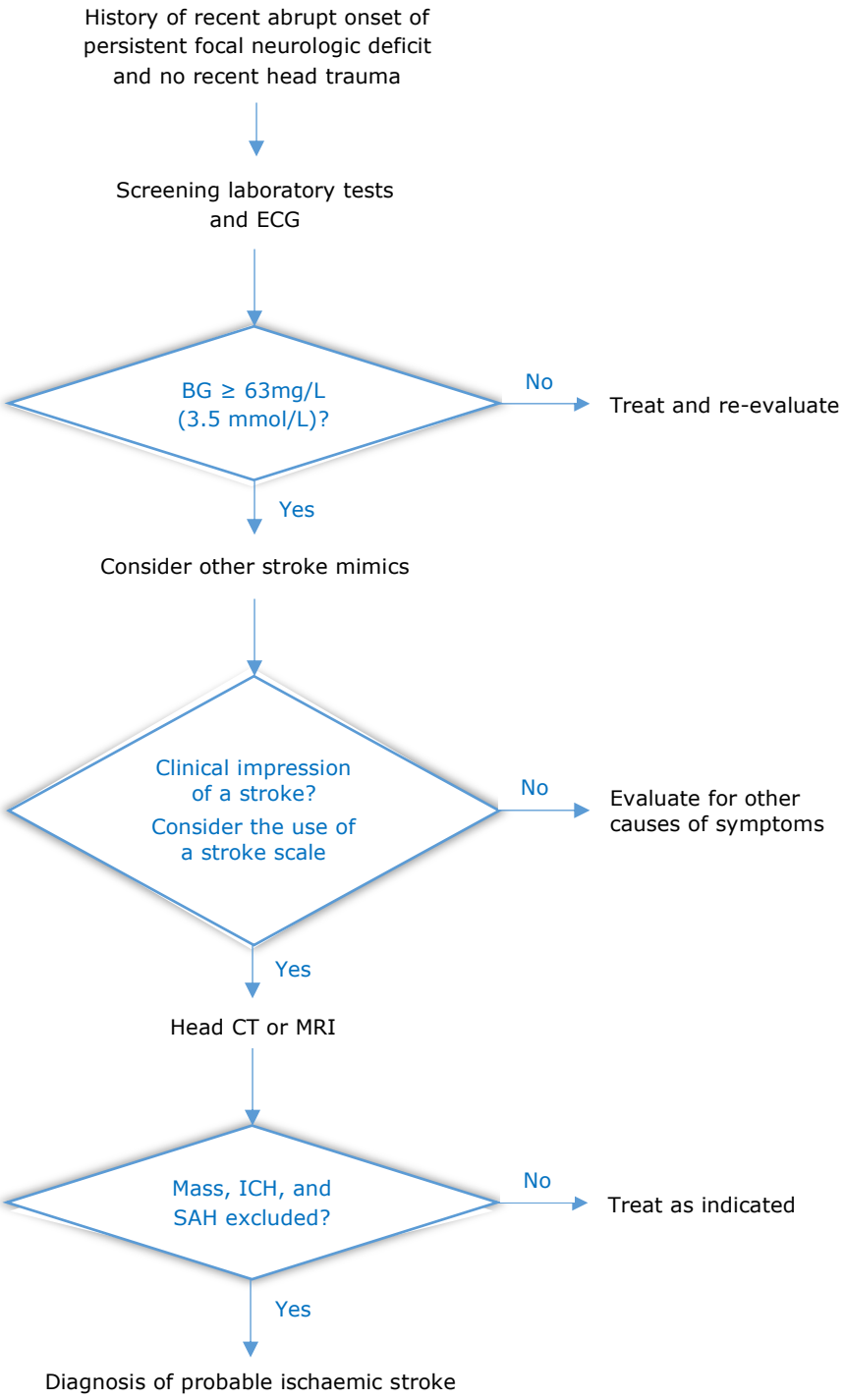


Figure 1.1 An example of a diagnostic algorithm for possible stroke

CT: computed tomography; ECG: electrocardiogram; ICH: intracerebral haemorrhage; MRI: magnetic resonance imaging; SAH: subarachnoid haemorrhage (Adapted from Yew, k., et al.[23])

1.2.3 Burden of stroke

Stroke is the second leading cause of death globally[25] and the fourth in the UK and accounts for significant disabilities in the same populations.[26,27] The epidemiology of stroke is changing over time, like any other non-communicable disease. Despite a decline in stroke mortality in many developed countries over the last few decades, globally stroke as a cause of death has moved from third to second place and is now the leading cause of physical disability in adults aged 65 years and older.[28] According to the WHO, about 15 million people have a stroke event each year. And of these, about a third (5 million) die and another 5 million are left with a permanent disability.[29] Figure 1.2 shows the variation in age-standardised stroke incidence rates between countries.

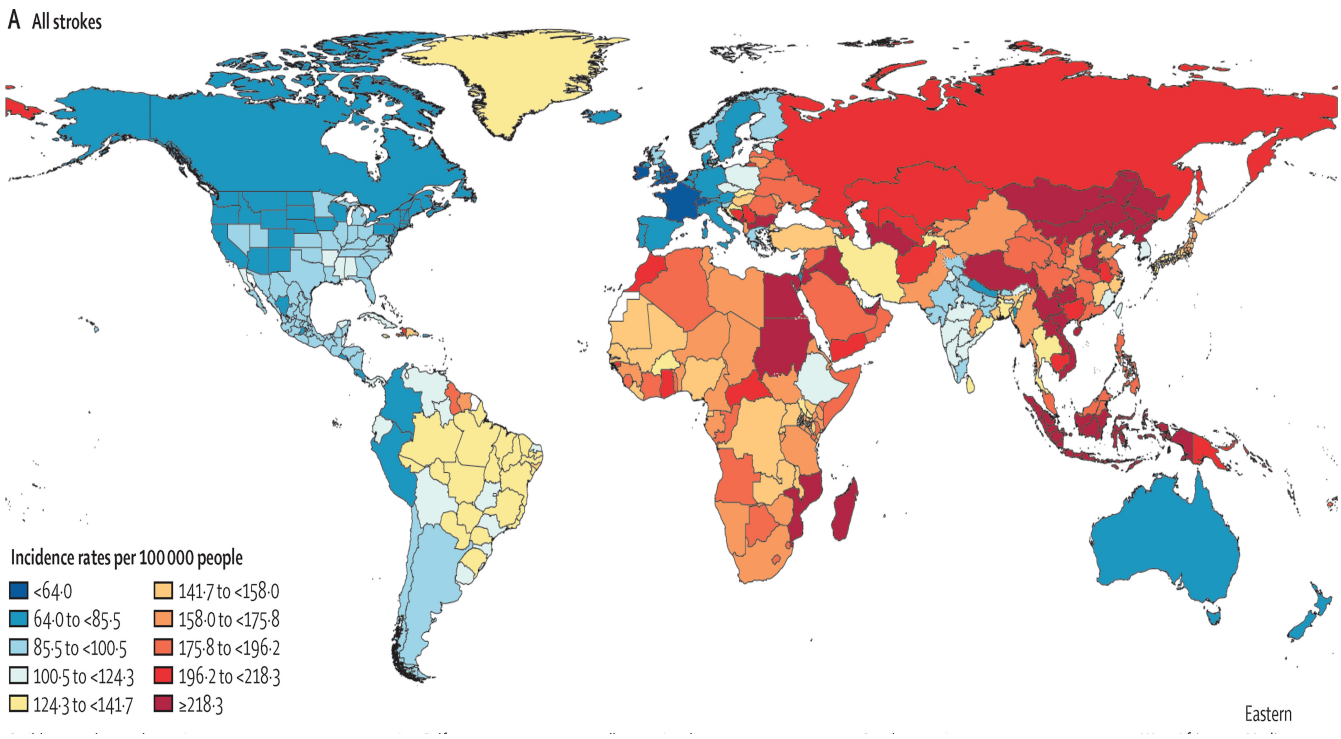


Figure 1.2 Age-standardised stroke incidence rates per 100,000 people for both sexes, 2019

All stroke types (haemorrhagic, ischaemic, and subarachnoid haemorrhage)

Adapted from Global Burden of Disease 2019 Stroke Collaboration [30]

There are over 1.3 million stroke survivors in the UK and about 100,000 people suffer strokes each year, with the average age at incident stroke being 73 years. The annual incidence of stroke cases in the UK is projected to increase by 60% between 2015 and 2035, whereas the number of stroke survivors is expected to more than double within the same period.[2]

1.2.3.1 Economic burden

Increases in the number of stroke survivors, advances in stroke treatment and rehabilitation, have resulted in increases in the financial burden posed by strokes. Using an individual patient simulation model, the total health and social care for patients with acute stroke each year in England, Wales, and Northern Ireland was £3.6 billion in the first 5 years after admission (mean per-patient cost: £46,039).[31] There was a fivefold variation in the magnitude of costs between patients, ranging from £19,101 to £107,336. The health and social care costs increased with older age (Figure 1.3), increasing stroke severity, and by having an intracerebral haemorrhagic stroke.[31]

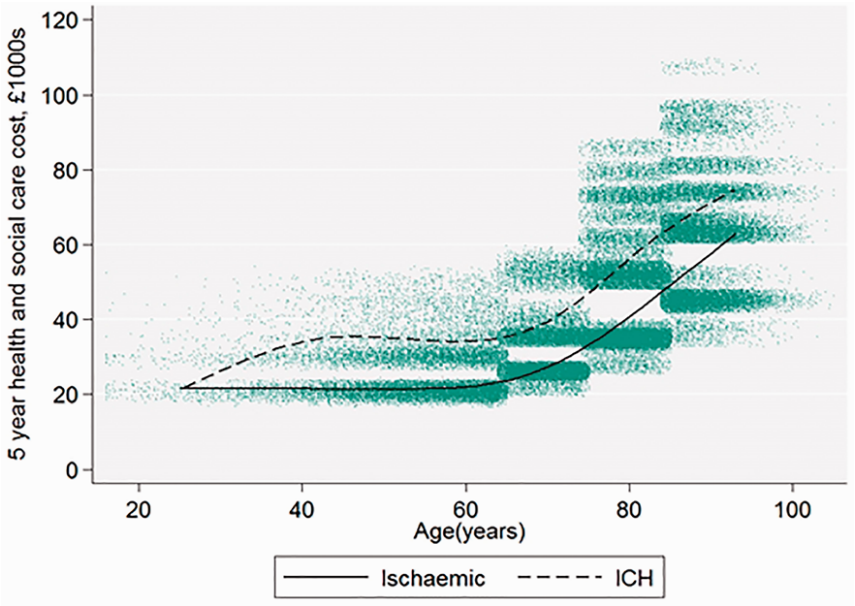


Figure 1.3 Five-year health and social costs by age and stroke severity for 84,184 patients admitted between April 2015 and March 2016

Each dot on the scatter plot is one patient with stroke; the best fit line is a restricted cubic spline with four knots.

Adapted from Xu X-M *et al.* [31]

1.2.4 Stroke risk factors

Risk factors for stroke may be grouped into biological (such as age and sex), physiological (such as hyperlipidaemia), behavioural (such as smoking and alcohol consumption), sociocultural (such as education and social class), and environmental (such as air pollution).[32] The heterogeneity of stroke makes it such that risk factors are often dependent on the specific type of stroke being dealt with. Thus, physical inactivity and cardiac factors may lead to ischaemic stroke, but the less frequent haemorrhagic stroke often has smoking as a major risk factor.[33]

Stroke risk factors can be sub-divided into modifiable and non-modifiable factors – Table 1.1.[33,34]

Table 1.1 Stroke risk factors

	Non-modifiable risk factors	Modifiable risk factors
Ischaemic stroke	Age	Hypertension
	Sex	Current smoking
	Ethnicity	Waist-to-hip ratio
	Apolipoprotein B to A1 ratio	Diet
		Physical inactivity
		Hyperlipidaemia
		Diabetes mellitus
		Alcohol consumption
		Cardiac causes
Haemorrhagic stroke	Age	Hypertension
	Sex	Current smoking
	Ethnicity	Waist-to-hip ratio
		Alcohol consumption
		Diet

Adapted from Boehme A.K. *et al*. [33] and O'Donnell M.J. *et al*. [34]

Modifiable Risk Factors

These are risk factors that may be potentially altered to prevent stroke or its sequalae. Hypertension is an essential risk factor due to its high prevalence in patients with stroke, and its favourable response to intervention strategies.[35] Tobacco use is also a dose-response risk factor which has an increased risk of stroke in heavy users as compared to lighter users. Obesity, physical inactivity, alcohol use, poor diet, and high blood cholesterol are all examples of modifiable risk factors for stroke.[36] These risks must be tackled through community-based interventions, that focus on environmental and lifestyle changes as preventive strategies against stroke.[17]

Non-modifiable Risk Factors

Some risk factors, including age, sex, and ethnicity, cannot be modified. Stroke incidence increases with increasing age in the general population, and the risk is estimated to double yearly after 55 years.[37] Studies have shown that strokes are

more common among women than men. The use of hormonal medications and hypertension during pregnancy have been linked to increased stroke rates among women.[38] New evidence suggests that the high incidence of strokes in women compared to men may also be attributable to other factors such as age and socioeconomic status. With regards to race and ethnicity, people of African origin are twice more likely to have a stroke episode compared to Caucasians and more likely to die from stroke or its associated complications.[33]

Heredity (genetic factors), generally considered to be non-modifiable, contribute to stroke.[33] With the availability of genetic therapeutics and modification of genetic factors through gene-environment interactions, some genetic factors may become modifiable in the future.[33]

The burden of stroke, therefore, can be reduced if risk factors are identified and understood, and efficacious measures are employed to reduce the risk, especially from modifiable risk factors.

1.2.5 Comorbidities associated with stroke

The occurrence and sequelae of most strokes are largely dependent on the presence of comorbid conditions.[39] Studies indicate that more than half of patients with stroke have hypertension, 20% diabetes, and about 13% have dyslipidaemia.[40,41] Prevalent comorbid conditions in the UK, in addition to the above, include coronary heart disease (19%), atrial fibrillation (18%), heart failure (7%), and peripheral vascular disease (4%).[42] Depression and dementia are increasingly being recognized as sequelae of stroke as well.[42,43] The mortality risk of stroke survivors also appears to increase with increasing comorbidities. Over 80% of stroke survivors also have at least one or more comorbid conditions.[44] The risk of mortality is doubled in stroke survivors with more than five comorbidities when compared to survivors with no comorbidity. And risk

factors such as age, smoking, and low socioeconomic status also increase the risk of mortality.[44]

1.2.6 Management of stroke

The primary aim for the acute management of patients with stroke is to within a short time frame, stabilise the patient, and complete the initial assessments which include imaging and laboratory tests. Four main areas in early stroke management have been demonstrated to reduce mortality and improve functional outcomes following stroke (Table 1.2):

- Early/rapid recognition of symptoms and diagnosis
- Emergency treatment of a patient with stroke
- Specialist care for the patient with acute stroke: patient-centred and goal-orientated care including early mobilisation where possible.

Over the last 2 decades, significant changes have been made in the care of patients with stroke, to improve the quality and efficiency of care provided. Key to this is the setting up of multidisciplinary stroke units comprising doctors, nurses, physiotherapists, dieticians, pharmacists, occupational therapists, speech and language therapists, clinical/neuropsychologists, and social workers.[45] The establishment of these stroke units and post-stroke rehabilitation programmes have been shown to reduce death and disability through the provision of specialised multidisciplinary care for diagnosis, emergency treatments, prevention of complications, rehabilitation and secondary prevention.[46,47]

Table 1.2 Acute management of patients with stroke

Content	Drivers	Interventions
Improve the outcomes for patients following a stroke	First Hour Bundle: Rapid recognition of symptoms and diagnosis within 3 hours	• Rapid diagnosis using recognised tools e.g., Recognition of Stroke in the Emergency Room (ROSIER) Scale • Confirmation of diagnosis by an experienced clinician • Start aspirin
	First Day Bundle: Emergency treatment for people with stroke within 24 hours	• CT scan • Admission to co-located beds • Swallow screen • Prescription of regular aspirin (if non-haemorrhagic stroke)
	First 3 Days Bundle: Early mobilisation following stroke within 3 days	• 36 hours continuous physiological monitoring • Manual handling assessment • Physiotherapy assessment commended • Getting patients out of bed
	First 7 Days Bundle: Patient-centred and goal-oriented specialist care following stroke within 7 days	• Occupational Therapy assessment recommended • MDT goal-setting meetings • Information sharing with patients/carers • Estimating discharge dates
	Reduce the number of episodes of avoidable harm	Interventions identified by the Global Trigger Tool analysis

Adapted from: NHS Wales Informatics Service. Health in Wales: Acute Stroke [48]

1.2.7 Prevention of stroke

Disease prevention relies on anticipatory actions that can be categorised as primary, secondary, or tertiary prevention.[49] The prevention of CVD can, therefore, occur at these three levels: primary (preventing or delaying the onset of disease); secondary (screening or early detection of disease); or tertiary (reducing the progression of the disease and managing disability and complications among people living with CVD). Goldston (1987),[50] however, notes that these levels might be better described as "prevention, treatment, and rehabilitation".

The term 'secondary prevention' can be used differently by different groups to refer to different aspects of preventive care. Secondary prevention consists of early diagnosis and prompt treatment to contain the disease and "disability limitation" to prevent potential future complications and disabilities from the disease.[49] This PhD research uses the definition common in clinical literature, where it simply refers to the prevention of recurrences or complications,[51] which overlaps with the concept of 'tertiary prevention' as defined in the fields of public health and epidemiology.

Tertiary prevention attempts to reduce the damage caused by symptomatic disease. The objective of tertiary prevention is to maximise the remaining capacity and functions of an already disabled patient.[52]

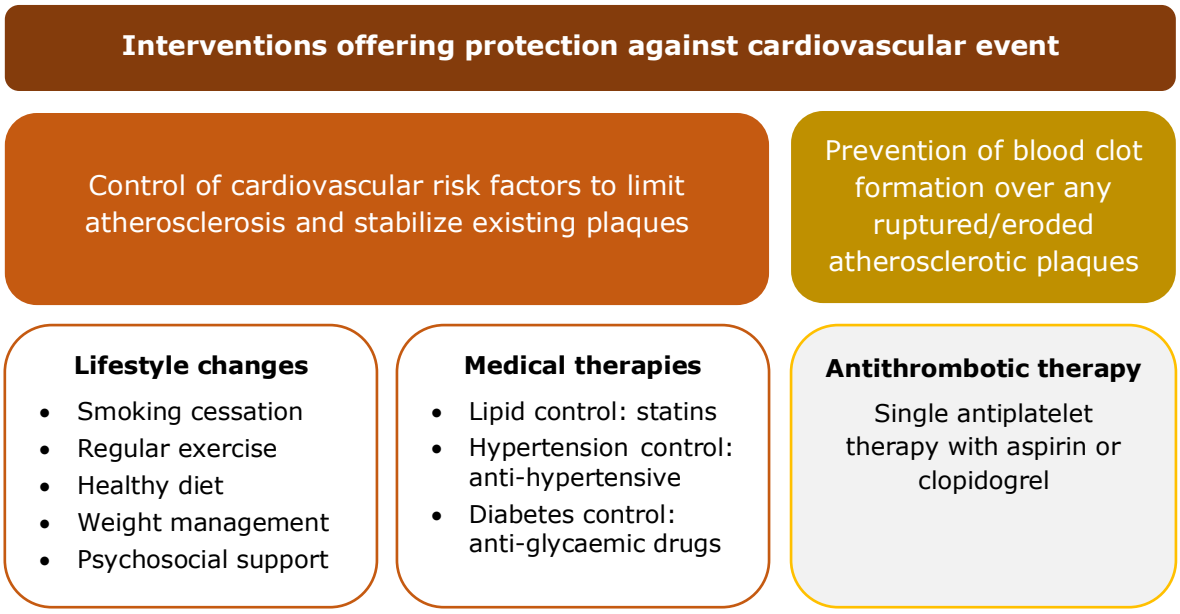


Figure 1.4 Non-pharmacological and pharmacological interventions to reduce the risk of cardiovascular disease

Adapted from: Thrombosis Advisor Resources [53]

1.2.7.1 Strategies for secondary prevention

Stroke survivors have a high risk for cardiovascular disease or another stroke, as well as for exacerbation of other conditions that led to the incident stroke.[17] Thus, a quick mnemonic for secondary stroke prevention emphasizes the use of the following[54]:

A – Antiplatelet and anticoagulant therapy

B – Blood pressure lowering medication (antihypertensives)

C – Cholesterol/lipid-lowering medication, cessation of smoking, and carotid revascularization

D – Dietary changes

E – Exercise or physical activity

The use of antiplatelets, antihypertensives, cholesterol-lowering medicine, and cessation of smoking have been shown to have the potential to reduce the risk of recurrent vascular outcomes by about 75%.[55] These strategies must therefore be applied on a population level as well as a patient-focused level to improve adherence and efficiency.[36]

Secondary preventive measures can be further classified into pharmacological and non-pharmacological strategies:

a. Non-pharmacological strategies

Despite sophisticated advances in medications aimed at reducing residual cardiovascular risk in patients with established stroke, harmful environments or exposures increase vascular risk.[56] These factors (such as low socioeconomic status, smoking, and making unhealthy food choices) negatively influence the prognosis of stroke.[36] Non-pharmacological interventions focused on changing behaviours are essential in reducing CVD risk.[57] These include:

- *Lifestyle advice**:*** intensive lifestyle advice is recommended to be given simultaneously with drug treatment.
    - *Smoking cessation* – stroke survivors are strongly encouraged to stop smoking and are supported in their efforts to do so. Cessation of other forms of tobacco use is also recommended and nicotine replacement therapy is recommended for patients who are likely to be markedly dependent on nicotine.
    - *Dietary changes* – The Dietary Approaches to Stop Hypertension (DASH) diet, has for instance been recommended to reduce the risk of recurrent stroke and/or CVD.[58] Reductions in the daily intake of salt (<5g per day), total fat (<30% of calories), saturated fat (<10% of calories), and eliminating trans-fatty acids are recommended. Patients are encouraged to eat a range of fruits and vegetables daily, whole grains, and use poly-saturated or mono-unsaturated dietary fats. Eating at least 5 portions of fruits and vegetables from a variety of sources each day.[59]
    - *Physical activity* – Regular light to moderate supervised physical exercise is recommended for patients recovering from a stroke. Engaging in at least moderate-intensity aerobic activity of at least 10 minutes 4 times a week or vigorous-intensity aerobic activity for at least 20 minutes twice a week has been shown to lower the risk of recurrent stroke and composite cardiovascular outcomes (recurrent stroke, myocardial infarction (MI), or vascular death).[60]
    - *Weight control* – In patients who are overweight or obese, weight loss is advised through the combination of a reduced energy diet and increased physical activity.

- *Alcohol intake* – Individuals who take more than 3 units of alcohol per day are advised to reduce alcohol consumption. Alcohol intake should be limited to 14 units per week, spread over at least 3 days.[59]

- *Stroke/cardiac rehabilitation* – is the coordinated system of care necessary to help people with stroke return to an active and satisfying life and helps to prevent the recurrence of stroke or new cardiovascular conditions. Effective rehabilitation services include the following components in addition to appropriate specialist medical care:
    - Individual assessment
    - Modification of risk factors
    - Purpose-designed exercise programmes
    - Health education and counselling
    - Behaviour modification strategies
    - Support for self-management

b. Pharmacological strategies

Current clinical guidelines [European Guidelines on Cardiovascular Disease Prevention in Clinical Practice,[61] The American Heart Association and the American College of Cardiology (AHA/ACC) Guidelines[62,63] and National Institute for Health and Care Excellence (NICE)[59,64]] recommend the use of pharmacological interventions unless contraindicated.

- *Antihypertensive treatment:* Blood pressure reduction is recommended in all patients with established CHD, TIA, or stroke, particularly with a blood pressure level above $^{140}/_{90}$ mmHg. A target blood pressure of $^{130}/_{80\text{-}85}$ mmHg is recommended.

- *Lipid-lowering treatment:* Lipid-lowering drugs are an essential pillar of secondary prevention. Life-long treatment with statins is recommended

in all patients with established CHD, stroke, or TIA. Intensive lipid-lowering strategies focused on reducing LDL-C (to under 70mg/dL or a reduction of at least 50% if the baseline LDL-C is between 70-135mg/dL) decreases the recurrence of cardiovascular events.[61,65,66] Ezetimibe or a proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitor (as monotherapy or as combined therapy) is recommended for patients whom it is not possible or is insufficient to achieve adequate cholesterol control with only statins.[61,67]

- *Antiplatelet treatment* with regular aspirin in the absence of clear contraindications or clopidogrel is recommended in patients with ischaemic stroke or permanent atrial fibrillation for long-term vascular prevention.
- *Anticoagulant treatment:* Long-term anticoagulation is recommended for patients with ischaemic stroke or TIA who are in atrial fibrillation once intracranial bleeding and other contraindications (such as uncontrolled hypertension) have been excluded.
- *Hyperglycaemic treatment:* Anti-diabetics are recommended for patients with persistent fasting blood glucose >6 mmol/l despite diet control.

Data from clinical trials have consistently proven the efficacy of pharmacological interventions with aspirin, statins, and blood pressure-lowering medications in reducing the risk of cardiovascular events and total mortality in the ever-growing pool of patients in secondary prevention.[68–70] However, adherence to medication in patients with established CVD is low.[71] The multi-factorial issue of non-adherence is, therefore, one of the major challenges in secondary prevention.[61]

Therapeutic resources that one might reserve for high-risk patients could include: lower target values for blood pressure and cholesterol, more intensive follow-up, high-cost therapies (e.g. PCSK9-inhibition) or therapies with a high risk of adverse events (e.g. bleeding risk in dual antiplatelet therapy).[61]

c. Revascularisation surgeries

The outcomes of stroke episodes largely depend on prompt management of symptoms and effective therapeutic methods. For acute ischaemic strokes in a patient who presents within 4.5 hours of the onset of symptoms, a key recommended therapy is intravenous thrombolysis with recombinant tissue plasminogen activator (rt-PA).[72] Endovascular revascularization is also being used for early recanalization of occluded vessels in acute ischaemic stroke.[73]

Stroke, dementia, chronic kidney disease, and other cardiovascular conditions confer risk for one another and also share risk factors such as risk factors such as increasing age, smoking, diabetes mellitus, hypertension, hyperlipidaemia, and the associated pathophysiology of small vessel disease.[74,75] To optimally manage the possible atherogenic effect of these comorbid conditions to reduce the risk of subsequent cardiovascular morbidity and mortality outcomes, both non-pharmacological (that is, lifestyle modification)[76,77] and pharmacological (antihypertensives for blood pressure management;[78] lipid-lowering medications such as statins for cholesterol management;[79] antidiabetics for blood sugar control;[76] and antiplatelets/anticoagulants to manage arrhythmia[80]) strategies need to be prioritised in line with clinical guidelines.[60] Frequent monitoring/reviews to ensure treatment targets are being met is important.[81]

The next sections detail the aim and objectives of this thesis and provide an outline of successive chapters.

1.3 Aim & Objectives

As aforementioned in Section 1.1, by interrogating large clinical data patients with a diagnosis of stroke can be stratified into different risk groups according to risk factors (clinical characteristics) to inform more targeted personalised preventive interventions.

1.3.1 Research question

Can distinct patient clusters with different risks for subsequent major adverse cardiovascular morbidity and mortality outcomes (i.e., coronary heart disease, recurrent stroke, peripheral vascular disease, heart failure, CVD-related mortality, and all-cause mortality) be identified using multidimensional phenotypic data routinely collected in clinical practice for adult patients with incident stroke diagnosis?

1.3.2 Aim

This PhD research aims to identify clinical phenotypes (that is, patient characteristics and distinct patient clusters) that correlate with subsequent major adverse cardiovascular event outcomes in adults with an incident stroke diagnosis.

1.3.3 Objectives

The specific objectives to achieve the research aim are:

1. To summarise the available evidence on prognostic models and assess their accuracy for predicting MACE outcomes in adults with an established stroke diagnosis.
2. To describe the age, sex, and socioeconomic differences in the rates of first non-fatal stroke and subsequent major adverse cardiovascular outcomes.

3. To compare the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke.

4. To examine the relationship between body mass index and MACE outcomes during long-term follow-up in patients with any subtype of incident stroke.

5. To explore heterogeneity in clinical characteristics of adult patients with incident stroke and cluster patients based on phenotypic characteristics, and to assess the independent association between phenotypic clusters and subsequent cardiovascular morbidity and mortality outcomes.

1.4 Structure of the thesis

The objectives are addressed in the remaining chapters as outlined below and illustrated in the research study map (Figure 1.5):

Chapter 2: Methods

The databases used in this research are described in this chapter. The study population (exposure), outcomes, and covariates are also defined. The chapter also provides an overview of the study methods used. However, detailed methods relevant to respective studies are provided in the corresponding chapter.

Chapters 3 – 7

Each chapter focuses on one of the afore-mentioned thesis objectives and also describes in detail the methods used for each respective study. Results for each study are presented and findings appropriately discussed.

- *Objective 1* is covered in Chapter 3 which provides a summary of the available evidence on prognostic models developed in adults with an established stroke diagnosis to predict subsequent MACE outcomes.

- *Objective 2* is covered in Chapter 4 which describes the age, sex, and socioeconomic differences in the rate of first non-fatal stroke and subsequent MACE outcomes.
- *Objective 3* is covered in Chapter 5 and compared the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and those with incident ischaemic stroke.
- *Objective 4* is covered in Chapter 6. Obesity is a risk factor for stroke and also a risk factor for hypertension and diabetes mellitus, which are risk factors for CVD.[82] Body mass index (BMI) is a common measure of obesity. The relationship between BMI and subsequent MACE outcomes in patients with incident stroke is assessed in this chapter.
- *Objective 5* is covered in Chapter 7 which explores multimorbidity clusters in patients with incident stroke. The chapter explores heterogeneity in clinical characteristics of adult patients with incident stroke by clustering patients based on phenotypic characteristics. The association between the identified phenotypic clusters and subsequent cardiovascular morbidity and mortality outcomes is further assessed.

Chapter 8: Summary conclusions and future directions for research

This final chapter summarises the main findings and discusses the clinical as well as public health implications. The chapter concludes with recommendations for further research studies.

The research chapters are reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist[83] for *chapter 3* and according to STrengthening the Reporting of OBservational studies in Epidemiology (STROBE)[84] for *chapters 4 – 7*.

Chapters	Objectives	Methods
Chapters 1 & 2 Introduction & Methods	To provide a general overview of what is already known about stroke, the rational for this work (aims/objectives), the datasets and the methods used.	
Chapter 3 Systematic review of stroke prognostic models	To summarise the available evidence on prognostic models and evaluate their accuracy for predicting major adverse cardiovascular event (MACE) outcomes in adult with established stroke diagnosis.	• Narrative synthesis
Chapter 4 Age, sex, and socioeconomic differences in stroke incidence and subsequent outcomes	To describe the age, sex, and socioeconomic differences in the rates of first non-fatal stroke and subsequent major adverse outcomes.	• Incidence rate ratio analysis
Chapter 5 Differences in subsequent outcomes: haemorrhagic vs. ischaemic stroke	To compare the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke.	• Propensity-score matching analysis • Cox proportional hazard regression
Chapter 6 Obesity paradigm in patients with incident stroke	To examine the relationship between body mass index and MACE outcomes during long-term follow-up in patients with any subtype of incident stroke.	• Cox proportional hazard regression
Chapter 7 Multimorbidity clusters in stroke	• To explore heterogeneity in clinical characteristics of adult patients with incident stroke and cluster patients based on phenotypic similarities. • To evaluate the association between the phenotypic clusters and occurrence of subsequent cardiovascular morbidity and mortality outcomes.	• Cluster analysis (unsupervised machine-learning approach) • Cox proportional hazard regression
Chapter 8 Summary conclusions and direction for further research	• To summarise the main findings and limitations for this thesis. • To make recommendation for further research.	

Figure 1.5 Research study map for thesis

Summary

This chapter has outlined the rationale for this thesis research, provided an overview of what is already known about stroke and its prevention and management, detailed the aim and objectives of the thesis, and provided an outline of successive chapters.

COVID-19 Impact Statement

In response to the coronavirus (COVID-19) pandemic and national guidelines and/or restrictions (national lockdowns), the University adjusted its ways of working and adapted remote working for the 2019/2020 and 2020/2021 academic years. In-person research activities were suspended. With the support of my supervisors, I resorted to remote working to progress my database work and with periodic supervision meetings held virtually.

Chapter 2

Methods

Chapter 1 described the research area of stroke, presented the research context, aims, and objectives for this thesis. This chapter focuses on the data used and outlines the relevant methodological approaches used in this thesis research. Further details of methodological approaches specific to each study have been reserved for the respective chapters. This chapter serves as a transition from discussing the thesis research context to focusing on the individual original studies.

2.1 Overview

This chapter initially describes the data sources used in addressing the research questions in this thesis. The data management principles applied to defining the study population, the outcome and covariate variables, are subsequently explained in detail. Finally, the last section will provide a brief overview of the statistical analysis methods used – systematic review and meta-analysis, Cox proportional hazards regression, propensity score matching, unsupervised machine learning approach (cluster analysis). Further details on methods specific to a study (research chapter), including statistical analyses, have been reserved for the respective chapters.

2.2 Data source

The research studies used linked electronic health records in the UK: primary care data from Clinical Practice Research Datalink (CPRD GOLD), secondary care data from Hospital Episodes Statistics Admitted Patient Care (HES APC) data, Office for National Statistics (ONS) mortality data, and social deprivation data.

2.2.1 Clinical Practice Research Datalink

Under the UK National Health Service (NHS), visits to the General Practitioner (GP) are free of any charge. Hence, over 98% of the population are registered with a primary care GP.[85] Each patient is assigned a unique NHS number and patient data are routinely recorded – providing good capture of primary care health information in a longitudinal electronic health record. Clinical Practice Research Datalink (CPRD) collects de-identified patient data (including diagnoses, symptoms, prescriptions, referrals, and tests) from a network of GP practices across the UK[86,87] and produces one of the largest ongoing databases of longitudinal medical records from primary care globally.[87] The CPRD database encompasses primary care records for over 35 million patients, with a median

follow-up time of 10 years, including 25% of patients with over 20 years of follow-up. CPRD collects data from practices using Vision® software that contribute to the CPRD GOLD database, which has been used in epidemiological research for 30 years.[87]

CPRD data is provided to researchers as different file types linked by a unique patient identifier. Table 2.1 provides information about the different file types.

Table 2.1 Summary of CPRD file types

CPRD Data File	Information contained
Patient	Demographics and registration details for the patients such as sex, year of birth, most recent date of joining the practice
Practice	Details of each practice, including the region where the practice is based, the date when practice data was deemed to be of research quality (i.e., up-to-standard date) and the date of last data collection
Staff	Practice staff details, with one record per member of staff
Consultation	Information relating to the type of consultation from a pre-defined list including a visit to the practice, telephone call, discharge summary or hospital letter.
Clinical	Has the medical history including symptoms, signs, and diagnosis of patients coded using Read codes. Diagnoses made during a hospital admission are also recorded based on hospital letters and discharge summaries.
Additional clinical details	Additional information linked to clinical information entered in the structured data area by the General Practitioner. Includes information such as smoking status, alcohol consumption, etc.
Referral	Referrals to external care centres (normally to secondary care locations such as hospitals for inpatient or outpatient care) and include speciality and referral type
Immunisation	Details of immunisation records including vaccinations offered to patients, vaccines accepted and whether immunisation is routine.
Test	Details relating to test data – the type of test, test results (either qualitative entries (e.g., normal or abnormal) or quantitative entries involving a numeric value
Therapy	Details of all prescriptions (for drugs and appliances) issued

The structure of the linked CPRD file types is shown in Figure 2.1.

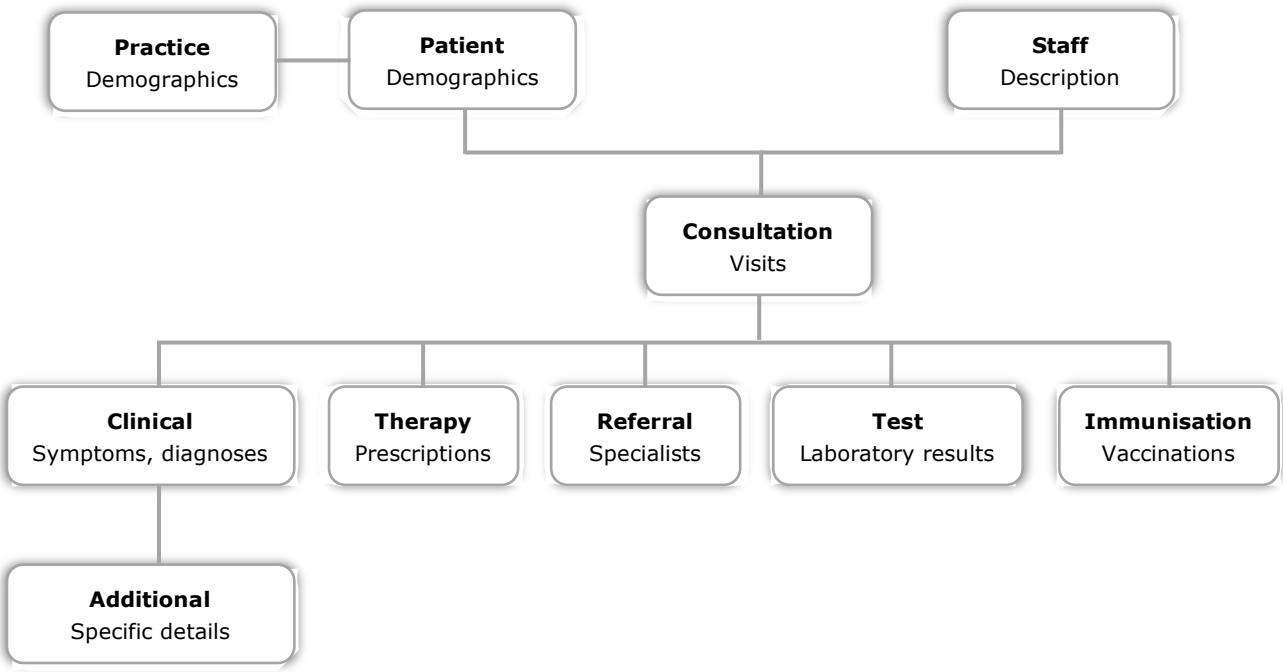


Figure 2.1 CPRD database structure

Emily Herrett et al., Int. J. Epidemiol. 2015; 44: 827-836

Patient-level data from about 58% of all general practices contributing to the CPRD GOLD database can be linked to other existing data sources through a trusted third party (NHS Digital).

2.2.1.1 Data quality

The CPRD data is routinely monitored by internal processes. CPRD has two key methods of ensuring high-quality research data is made available to the researchers: the 'acceptable research quality' flag (Appendix A), which is a patient-level quality marker; and the 'up-to-standard date' (Appendix A), a practice-level quality marker.[88]

The quality of the data recorded has also been improved by the primary care pay-for-performance scheme introduced in 2004 – the Quality and Outcome Framework (QOF).[89] The documentation/coding for specific disease conditions and associated risk factors have significantly improved as a result of QOF.

2.2.1.2 Generalisability and validity of the data

Patients in CPRD GOLD are broadly representative of the UK general population in terms of sex, age, ethnicity.[87,90] CPRD is very widely used internationally for epidemiological research and has been used to produce over 2,000 research studies published in peer-reviewed journals across a broad range of health outcomes.[87]

2.2.2 Hospital Episodes Statistics

Hospital Episodes Statistics Admitted Patient Care (HES APC) data is a national data set of all admissions to National Health Service (NHS) hospitals in England.[91] This includes admissions to private or charitable hospitals paid for by the NHS. Almost all hospital activities (98 – 99%) in England are funded by the NHS.[92] Data related to accident and emergency (A&E, emergency department) attendance or outpatient hospital appointments are held in separate HES databases collated and curated by NHS Digital.[91] Due to its universal coverage, long period of data collection, and the ability to follow individuals over time, HES APC has been extensively used for research and health service evaluation.

In 1997/98 NHS numbers became a mandatory return from hospitals. As a result, HES APC is routinely linked to external datasets including the primary care database, CPRD GOLD. In addition to data on diagnoses and procedures, HES APC contains information on dates of admission, operations, and discharge. Diagnoses are recorded using the International Statistical Classification of Diseases and Health-Related Problems, 10th revision (ICD-10).[93] Operations and other

interventions are coded using a UK-specific coding system, the Office of Population Censuses and Surveys Classification of Interventions and Procedures (OPCS).[94]

2.2.3 Office of National Statistics Mortality Data

The Office of National Statistics' (ONS) mortality data provides data on the place of death and the original underlying cause of death which takes into account information provided by medical practitioners and/or coroners.[95,96] Linking ONS mortality data to HES data permits the analysis of deaths in and outside the hospital for all patients with a record in HES. Cause-specific mortality data are recorded using ICD-10.

2.2.4 Index of Multiple Deprivation

Socioeconomic status (SES) is measured by the English Index of Multiple Deprivation (IMD) 2015[97] which is a widely used composite measure of small-area based deprivation. The current IMD measure quantifies relative deprivation across seven different domains of deprivation: income; employment; education, skills and training; health and disability; crime; barriers to housing and services; and living environment, using the Lower Super Output Area (LSOA) level, where an LSOA contains between 400 and 1200 households.[98] The overall IMD is calculated as a weighted mean across the seven domains, hence offering a single score to describe the concept of deprivation whiles recognising the many interacting components. SES is ranked into quintiles (quintile 1 – highest SES group to quintile 5 – lowest SES group). Individual-level measures of SES such as educational level or income are not available.

2.2.5 Data anonymity and ethical approval

All linked patient electronic health record is anonymised to ensure patients cannot be identified by researchers. Each patient has a unique identification number to enable the linkage of the various datasets. This study was approved by the

Independent Scientific Advisory Committee for the Medicines and Healthcare Products Regulatory Agency (ISAC Protocol 19_023R) – see Appendix 1.

2.3 Data management

The 2019_09 database build for CPRD GOLD was obtained and used for this thesis research. The raw data files were provided as text flat files by the CPRD fob holder for the University of Nottingham. HES APC, mortality and social deprivation (IMD) data were provided by CPRD for the patient cohort eligible for linkage. Data management of the entire datasets was done by the doctoral researcher.

2.3.1 Study population

A cohort of patients with the first record of any non-fatal stroke in either CPRD GOLD or HES APC data between 1 January 1998 and 31 December 2017 were identified, Figure 2.2. Patients entered the study cohort at the minimum date of study start (1 January 1998); being aged 18 years and over; with at least 12 months of registration;[99] practice data of acceptable quality ('up-to standard'); and eligible for linkage to HES.

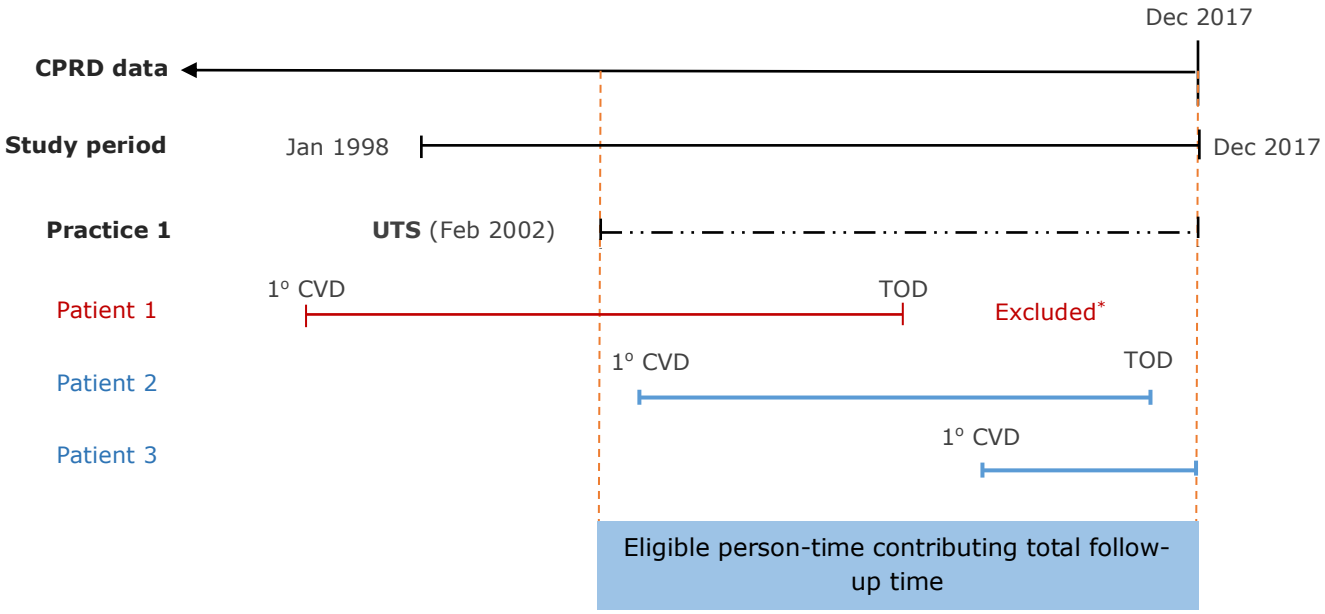


Figure 2.2 Schematic diagram showing patient eligibility and follow-up time

TOD: Transfer out date; *UTS*: up-to-standard date; *1° CVD*: incident cardiovascular event

The follow-up end date was defined as the date of transfer out/leaving the practice, date of death, last date for CPRD GOLD – HES APC link, or the last date of data collection, whichever came earliest. Patients with a prior history of stroke before the study start date (1 January 1998) were excluded from the study.

To assess the incidence of subsequent major adverse outcomes, the cohort consisted of patients with no prior history of major adverse events (that is, coronary heart disease (CHD) including coronary revascularisation, peripheral vascular disease (PVD) or heart failure). Figure 2.3 shows the study flow diagram.

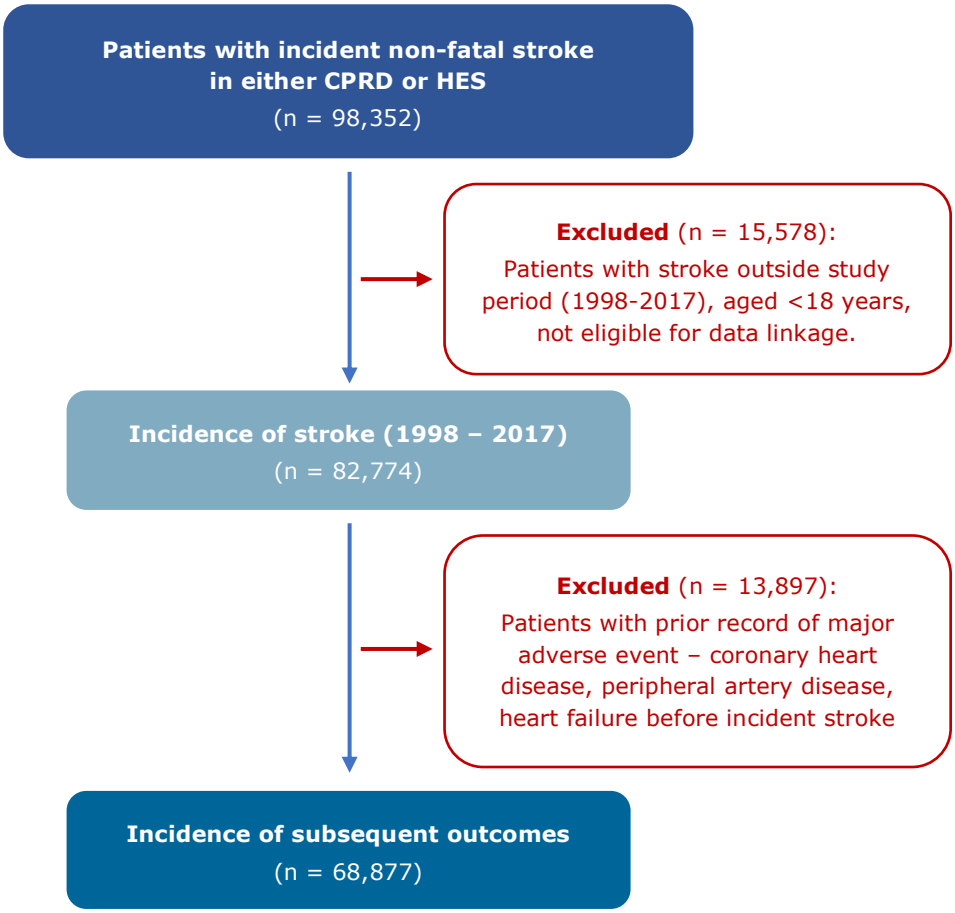


Figure 2.3 Study flow diagram

2.3.2 Study variables

This section describes and details the variables used throughout this PhD research. Variables (covariates) used for each study will be detailed in the respective chapter (chapters 4–6).

2.3.2.1 Definition of stroke

The focus of this research thesis is on stroke and for the database studies, patients with an incident diagnosis of non-fatal stroke. To identify patients with a coded diagnosis of stroke, Read codes for primary care CPRD GOLD dataset (Table 2.2) and International Classification of Diseases, tenth revision (ICD-10) codes for secondary care HES APC (Table 2.3) were used.

Table 2.2 Read codes for stroke

Read code	Description
Haemorrhagic stroke	
662o.00	Haemorrhagic stroke monitoring
7004300	Evacuation of intracerebral haematoma NEC
G610.00	Cortical haemorrhage
G611.00	Internal capsule haemorrhage
G612.00	Basal nucleus haemorrhage
G613.00	Cerebellar haemorrhage
G614.00	Pontine haemorrhage
G615.00	Bulbar haemorrhage
G616.00	External capsule haemorrhage
G617.00	Intracerebral haemorrhage, intraventricular
G618.00	Intracerebral haemorrhage, multiple localized
G619.00	Lobar cerebral haemorrhage
G61..00	Intracerebral haemorrhage
G61..11	CVA - cerebrovascular accident due to intracerebral haemorrhage
G61..12	Stroke due to intracerebral haemorrhage
G61X000	Left-sided intracerebral haemorrhage, unspecified
G61X100	Right-sided intracerebral haemorrhage, unspecified
G61X.00	Intracerebral haemorrhage in hemisphere, unspecified
G61z.00	Intracerebral haemorrhage NOS
G681.00	Sequelae of intracerebral haemorrhage
G682.00	Sequelae of other nontraumatic intracranial haemorrhages
Gyu6200	Other intracerebral haemorrhages
Gyu6F00	Intracerebral haemorrhage in hemisphere, unspecified
Ischaemic stroke	
G63..11	Infarction - precerebral
G63y000	Cerebral infarct due to thrombosis of precerebral arteries
G63y100	Cerebral infarction due to embolism of precerebral arteries
G640000	Cerebral infarction due to thrombosis of cerebral arteries
G640.00	Cerebral thrombosis
G641000	Cerebral infarction due to embolism of cerebral arteries
G641.00	Cerebral embolism
G641.11	Cerebral embolus
G64..00	Cerebral arterial occlusion
G64..11	CVA - cerebral artery occlusion
G64..12	Infarction - cerebral
G64..13	Stroke due to cerebral arterial occlusion

Read code	Description
G64z000	Brainstem infarction
G64z100	Wallenberg syndrome
G64z111	Lateral medullary syndrome
G64z200	Left-sided cerebral infarction
G64z300	Right-sided cerebral infarction
G64z400	Infarction of basal ganglia
G64z.00	Cerebral infarction NOS
G64z.11	Brainstem infarction NOS
G64z.12	Cerebellar infarction
G683.00	Sequelae of cerebral infarction
G6W..00	Cerebral infarct due to unspecified occlusion/stenosis precerebral arteries
G6X..00	Cerebral infarction due/unspecified occlusion or stenosis/cerebral arteries
Gyu6300	Cerebral infarction due/unspecified occlusion or stenosis/cerebral arteries
Gyu6400	Other cerebral infarction
Gyu6G00	Cerebral infarct due to unspecified occlusion/stenosis precerebral arteries
Stroke Not-Otherwise Specified (NOS)	
14A7.00	H/O: CVA/stroke
14A7.11	H/O: CVA
14A7.12	H/O: stroke
14AK.00	H/O: Stroke in last year
1M4..00	Central post-stroke pain
661M700	Stroke self-management plan agreed
661N700	Stroke self-management plan review
662e.00	Stroke/CVA annual review
662e.11	Stroke annual review
662M100	Stroke 6-month review
662M200	Stroke initial post discharge review
662M.00	Stroke monitoring
7P24200	Delivery of rehabilitation for stroke
8HHM.00	Ref to multidisciplinary stroke function improvement service
8IEC.00	Ref multidisciplinary stroke function improvement declined
9h21.00	Excepted from stroke quality indicators: Patient unsuitable
9h22.00	Excepted from stroke quality indicators: Informed dissent
9h2..00	Exception reporting: stroke quality indicators
Fyu5600	Other lacunar syndromes
G663.00	Brain stem stroke syndrome
G664.00	Cerebellar stroke syndrome
G665.00	Pure motor lacunar syndrome
G666.00	Pure sensory lacunar syndrome

Read code	Description
G667.00	Left sided CVA
G668.00	Right sided CVA
G66..00	Stroke and cerebrovascular accident unspecified
G66..11	CVA unspecified
G66..12	Stroke unspecified
G66..13	CVA - Cerebrovascular accident unspecified
G68X.00	Sequelae of stroke, not specified as haemorrhage or infarction
Gyu6C00	Sequelae of stroke, not specified as haemorrhage or infarction
L440.11	CVA - cerebrovascular accident in the puerperium
L440.12	Stroke in the puerperium
ZV12511	Personal history of stroke
ZV12512	Personal history of cerebrovascular accident (CVA)

Table 2.3 ICD-10 codes for stroke

Haemorrhagic stroke	I61, I62.9, I69.1, I69.2
Ischaemic stroke	I63.0, I63.4, I63.9, I63.1, I63.5, I69.3, I63.2, I63.6, I63.3, I63.8
Stroke Not-Otherwise Specified (NOS)	I64, I69.4, I69.8, G46.3, G46.7, G46.6, G46.5, G46.4, G46.8

2.3.2.2 Outcomes

A five-component major adverse cardiovascular event (MACE) endpoint was used as the primary outcome. MACE was defined as a composite of CHD, recurrent stroke, PVD, heart failure, and CVD-related mortality based on records from CPRD GOLD, HES APC and ONS mortality data. All-cause mortality was a secondary outcome.

2.3.2.3 Covariates (Features)

Based on availability in the electronic health records and established association with CVD or MACE, 336 candidate variables were selected. Table 2.4 details the phenotype domains and individual phenotype variables that served as phenotypic features for cluster analysis in Chapter 7. Some of these covariates were also used as confounders in other studies – chapters 4–6. The phenotypic domains include

demographics, physical characteristics, vital signs, laboratory data, co-morbid conditions, and prescribed medications.

For vital signs and laboratory test results, the most recent values/records within 24 months before incident stroke were extracted. A prescription within 12 months before the incident stroke was considered as a medication prescribed. All comorbid conditions were defined based on the latest record of a comorbid condition any time before the incident stroke.

Table 2.4 Domains and individual variables

Domains	Variables
Demographics	Age at incident stroke, alcohol status, ethnicity, incident stroke subtype, index of multiple deprivations, sex, smoking status, year of incident stroke
Physical characteristics	Body mass index, diastolic blood pressure, height, pulse rate, systolic blood pressure, weight
Laboratory	Alanine aminotransferase (ALT), albumin, alkaline phosphatase (ALP), bilirubin, calcium (adjusted), calcium, creatinine, c-reactive protein, eosinophil, erythrocyte sedimentation rate (ESR), gamma-glutamyl transferase (GGT), glomerular filtration rate (GFR), glycated haemoglobin (HbA1c), haemoglobin (Hb), high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, HDL-LDL ratio, lymphocyte, neutrophil, platelet, potassium, sodium, thyroid-stimulating hormone (TSH), total cholesterol, triglyceride (TG), urea
Comorbid conditions – cancers	Benign neoplasm – brain, colon, ovary, stomach, uterus; haemangioma, leiomyoma, cancer (composite), Hodgkin lymphoma, leukaemia, metastatic tumour, monoclonal gammopathy of uncertain significance, myelodysplastic syndrome, non-Hodgkin lymphoma, non-metastatic cancer, plasma cell malignancy, polycythaemia vera, primary malignancy (biliary, bladder, bone, bowel, brain, breast, cervical, kidney, liver, lung, melanoma, oesophageal, oropharyngeal, other, ovarian, pancreas, prostate, skin, stomach, testis, thyroid, uterus), secondary malignancy (bone, brain, liver, lung, lymph nodes, others)

Domains	Variables
Comorbid conditions – circulatory system	Abdominal aortic aneurysm, arrhythmia, atrial fibrillation, atrioventricular block (first, second, and third-degree), cardiomyopathy – other, congenital septal defect, dilated cardiomyopathy, family history of cardiovascular disease, family history of coronary heart disease, hypertension, hypertrophic cardiomyopathy, left bundle branch block, multiple valve disorder, non-rheumatic aortic valve disorder, non-rheumatic mitral valve disorder, pericardial effusion, primary pulmonary hypertension, Raynaud's disease, rheumatic valve disorder, right bundle branch block, sick sinus syndrome, subarachnoid haemorrhage, subdural haematoma, supraventricular tachycardia, transient ischaemic attack, venous thrombosis (excluding PR), ventricular tachycardia
Comorbid conditions – digestive system	Alcoholic liver disease, autoimmune liver disease, Barrett's Oesophagus, cholangitis, cholecystitis, cholelithiasis, cirrhosis, coeliac disease, Crohn's disease, diverticular disease, fatty liver, gastritis and duodenitis, gastroesophageal reflux disease, irritable bowel syndrome, liver failure, mild liver disease, moderate-severe liver disease, pancreatitis, peptic ulcer disease, peritonitis, portal hypertension, ulcerative colitis
Comorbid conditions – diseases of the ear	Hearing loss, Meniere's disease, otitis media, tinnitus
Comorbid conditions – endocrine system	Cystic fibrosis, diabetes mellitus (composite), diabetes mellitus (type 1, type 2, with complications, and with no complications), dyslipidaemia, family history of hyperlipidaemia, hyperparathyroidism, hypoglycaemia-causing disorders, hypothyroidism, obesity, polycystic ovarian syndrome, thyroid disease *(hypo- or hyperthyroidism)*
Comorbid conditions – diseases of the eye	Anterior uveitis, blindness, cataract, diabetic ophthalmic complications, glaucoma, keratitis, macular degeneration, posterior uveitis, retinal detachment, retinal vascular occlusion, scleritis
Comorbid conditions – genitourinary system	Acute kidney injury, benign prostatic hyperplasia, chronic kidney disease, end-stage renal disease, erectile dysfunction, female infertility, glomerulonephritis, male infertility, neuropathic bladder, obstructive and reflux uropathy, proteinuria, renal disease, urinary incontinence, urolithiasis

Domains	Variables
Comorbid conditions – respiratory system	Allergic and chronic rhinitis, asbestosis, asthma, bronchiectasis, chronic obstructive pulmonary disease, chronic sinusitis, pleural effusion, pleural plaque, pneumothorax, pulmonary collapse, pulmonary fibrosis, respiratory failure, sleep apnoea
Comorbid conditions – haematological and immunological conditions	Agranulocytosis, anaemia – other, aplastic anaemia, folate deficiency anaemia, hypersplenism, hyposplenism, immunodeficiency, iron deficiency anaemia, other haemolytic anaemia, primary thrombocytopaenia, sarcoidosis, secondary polycythaemia, secondary thrombocytopaenia, sickle cell trait, thalassaemia, thalassaemia trait, thrombophilia, vitamin B12 deficiency anaemia
Comorbid conditions – infectious disease	Chronic viral hepatitis, HIV, rheumatic fever, tuberculosis
Comorbid conditions – mental health disorders	Alcohol misuse, anxiety, autism, bipolar affective disorder, conduct disorder, delirium, dementia, depression, eating disorders, hyperkinetic disorders, intellectual disability, insomnia, obsessive-compulsive disorder, personality disorder, schizophrenia, self-harm, severe mental illness, substance misuse
Comorbid conditions – musculoskeletal system	Ankylosing spondylitis, back pain, carpal tunnel syndrome, collapsed vertebra, connective tissue disease, enthesopathy and synovial disorder, fibromatosis, giant cell arteritis, gout, intervertebral disc disorder, lupus erythematosus, osteoarthritis, osteoporosis, polymyalgia rheumatica, psoriatic arthritis, reactive arthritis, rheumatoid arthritis, scleroderma, scoliosis, Sjogren syndrome, spinal stenosis, spondylolisthesis, spondylosis
Comorbid conditions – neurological conditions	Autonomic neuropathy, Bell's palsy, cerebral palsy, chronic fatigue syndrome, diabetic neuropathy, epilepsy, essential tremor, hemiplegia, migraine, motor neurone disease, multiple sclerosis, myasthenia gravis, Parkinson's disease, peripheral neuropathy, trigeminal neuralgia
Comorbid conditions – skin conditions	Acne, actinic keratosis, alopecia areata, dermatitis, hidradenitis suppurativa, lichen planus, pilonidal cyst/sinus, psoriasis, rosacea, seborrheic dermatitis, urticaria, vitiligo

Domains	Variables
Prescribed medications	Acarbose, angiotensin-converting enzyme inhibitor, alpha-blocker, antihypertensive, antiarrhythmic, anticoagulant, antidepressant, antidiabetic, antiepileptic, antiplatelet, anxiolytic, beta-blocker, bile acid sequestrant, calcium channel blocker, centrally acting antihypertensive, corticosteroid, diuretic, DPP-4 inhibitors (Gliptins), fibrates, Glinide, Glucagon-like peptide-1 (GLP-1), hormone replacement therapy, immunosuppressant, inotrope, loop diuretic, metformin, nicotinic acid, nitrates, non-steroidal anti-inflammatory drugs, opioid, oral contraception, peripheral vasodilator, proton pump inhibitor, RAAS inhibitor, sodium-glucose co-transporter-2 inhibitors, statin potency[2], sulfonylureas, thiazide diuretics, thiazolidinediones, vasodilator, warfarin

All code lists used have been published and available for download from Cardiovascular Disease Research using Linked Bespoke Studies and Electronic Health Records (CALIBER) - https://www.caliberresearch.org/portal and CPRD @ Cambridge, Primary Care Unit, University of Cambridge.[100,101]

2.3.3 Dealing with outliers

Outliers are data points that are distant from the remaining data cluster.[102] In general, outliers are a result of procedural errors such as inaccurate data collection/entry or are inherited from the natural variations within the study population or process. For example, an anomalous blood test result may be explained by severe physiological state, intake of medication(s), food, or alcohol or poor handling of a blood sample. Outliers could increase error variance, reduce statistical power, decrease normality in situations where the outliers are non-randomly distributed, or introduce bias in the true relationship between exposure

[2] Statins were grouped into 3 different intensity categories according to the percentage reduction in low-density lipoprotein level: *low intensity* (a 20 – 30% reduction); *medium intensity* (31 – 40% reduction); *high intensity* (more than 40% reduction).[308,309]

and outcome.[103] There are many different methods to detect outliers and these include visual inspection using plots, clustering, and local outlier factor.[104] Visual inspection using plots and summary statistics (determining the minimum and maximum values for variables) was done to assess for outliers.

Outliers are generally handled in one of three methods: retaining outliers like every other data point; trimming outliers (i.e., removing outliers); and winsorising outliers (converting outliers to the value of the highest data point not considered to be an outlier).[105] Retaining outliers may lead to estimates that significantly vary from the legitimate population value. The trimming is usually recommended for outliers likely to be due to typographical mistakes or measurement errors as outliers could be legitimate observations and indicate the natural variance of the data. In consultation with clinicians (2 General Practitioners and 2 Cardiologists), lower and upper-value limits were agreed for study variables. The trimming method was used to deal with outliers in this research thesis.

2.3.4 Dealing with missing values

Missing data are common in both epidemiological and clinical research using routinely collected data. Missing data may lead to bias and loss of information which could undermine the validity of epidemiological and clinical research results. The risk of bias due to missing data depends on the reason for the missing data. The reasons for missing data are usually classified as missing completely at

random (MCAR)[3], missing at random (MAR)[4], or missing not at random (MNAR)[5].[106]

In general, missing data is usually addressed using two main approaches: complete case analysis of individuals with no missing data in any of the variables required for the analysis or imputation method. A multivariable complete-case analysis excludes patients with missing values and this approach may reduce statistical power, precision, and introduce bias.[107]

It is not possible, however, to distinguish between MAR and MNAR using observed data. Bias introduced by data that are MNAR can only be addressed by sensitivity analyses examining the effect of various assumptions about the missing data mechanism. Analysis based on complete cases when data is MAR (although not completely at random) may yield biased results. Such bias, however, can be overcome using multiple imputation.[107]

Multiple imputation is a general-purpose, relatively flexible, and computationally intensive approach used in dealing with missing data. The proportion of missing values for the respective variables for this research study are presented in Figure 2.4. To estimate missing values, multiple imputation by chained equations (predictive mean matching) was used to generate 40 imputed datasets using all the other available patient variables with complete data and all the outcomes,[108–110] for this PhD research. The imputed datasets were pooled into a single dataset using Rubin's rules.[111]

---

[3] Missing completely at random (MCAR): there are no systematic differences between the missing values and the observed values.

[4] Missing at random (MAR): Systematic difference between the missing values and the observed values can be explained by differences in observed data.

[5] Missing not at random (MNAR): After taking the observed data into account, systematic differences between the missing and observed values remain.

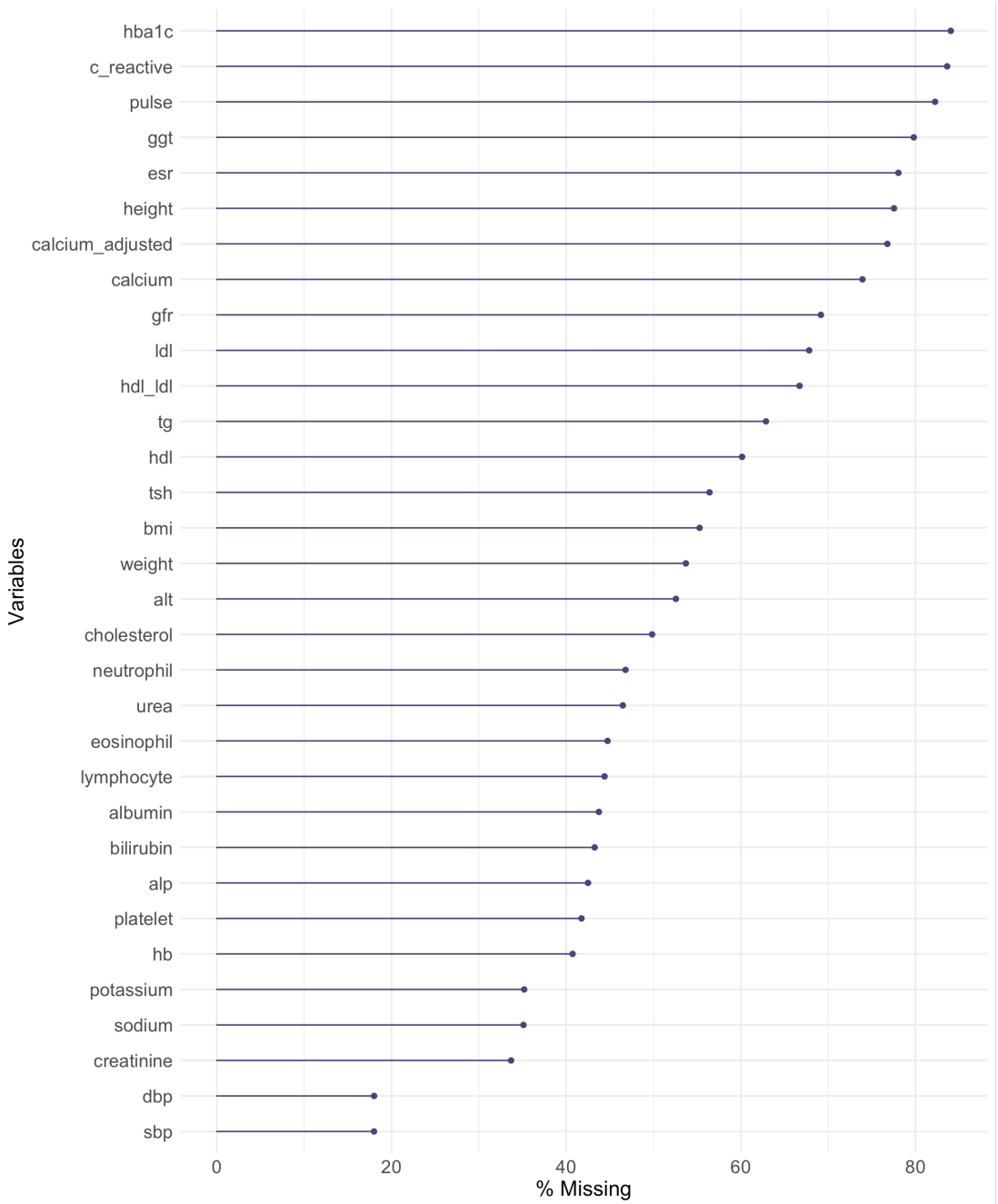


Figure 2.4 Proportion of missing values for the respective variables

The median with interquartile range (IQR) for both observed and imputed values for all the variables with missing values is presented in Table 2.5. Additionally, the imputed values from the multiple imputation method were inspected by comparing the distribution of the imputed and observed values – Figures 2.5a to 2.5e.

Table 2.5 Observed versus imputed values after multiple imputation for all clinical variables with missing data

Variables	Median (Interquartile range)	
	Observed	Imputed
Alanine aminotransferase	19.0 (15.0 – 27.0)	23.2 (21.23 – 25.83)
Albumin level	41.0 (38.0 – 43.0)	40.6 (39.9 – 41.2)
Alkaline phosphatase	82.0 (67.0 – 104.0)	95.0 (89.1 – 102.5)
Bilirubin level	10.0 (7.0 – 13.0)	10.9 (10.1 – 11.8)
Body mass index	26.3 (23.1 – 30.0)	26.5 (25.6 – 27.4)
Diastolic blood pressure	80 (71 – 85)	80 (78 – 81)
Systolic blood pressure	140 (130 – 150)	141 (139 – 143)
Calcium level (adjusted)	2.34 (2.27 – 2.41)	2.34 (2.32 – 2.36)
Calcium level	2.33 (2.26 – 2.41)	2.34 (2.32 – 2.36)
Creatinine level	87.0 (74.0 – 104.0)	92.2 (88.6 – 97.0)
C-reactive protein	5.0 (3.0 – 11.0)	10.7 (7.3 – 15.3)
Eosinophil level	0.2 (0.2 – 0.3)	0.3 (0.2 – 0.4)
Erythrocyte sedimentation rate	14.0 (7.0 – 27.0)	18.5 (14.3 – 22.7)
Gamma glutamyl transpeptidase	29.0 (19.0 – 51.0)	44.9 (36 – 58.5)
Glomerular filtration rate	66.0 (56.0 – 81.0)	67.2 (64.0 – 70.5)
Haemoglobin level	13.5 (12.4 – 14.6)	13.5 (13.2 – 13.9)
HbA1c level	47.5 (40.0 – 59.6)	50.1 (47.4 – 53.1)
HDL/LDL ratio	3.5 (2.0 – 4.4)	3.7 (3.4 – 4.0)
Height	1.65 (1.58 – 1.73)	1.67 (1.64 – 1.69)
HDL cholesterol	1.4 (1.1 – 1.7)	1.5 (1.4 – 1.6)
LDL cholesterol	2.9 (2.2 – 3.6)	3.0 (2.8 – 3.2)
Lymphocyte count	1.8 (1.4 – 2.4)	3.2 (2.6 – 3.9)
Neutrophil count	4.3 (3.4 – 5.6)	4.9 (4.6 – 5.6)
Platelet count	248.0 (200.0 – 302.0)	248.0 (234.4 – 261.7)
Potassium level	4.4 (4.1 – 4.7)	4.4 (4.3 – 4.5)
Pulse	76 (68 – 84)	76 (74 – 79)
Sodium level	140 (137 – 142)	139 (138 – 140)
TSH level	1.8 (1.2 – 2.7)	2.1 (2.0 – 2.3)
Total cholesterol level	5.0 (4.2 – 5.8)	5.1 (4.9 – 5.3)
Triglyceride level	1.3 (1.0 – 1.8)	1.5 (1.3 – 1.6)
Urea	6.0 (4.8 – 7.6)	6.4 (5.9 – 6.9)
Weight	73.0 (61.6 – 85.0)	74.2 (70.8 – 77.6)

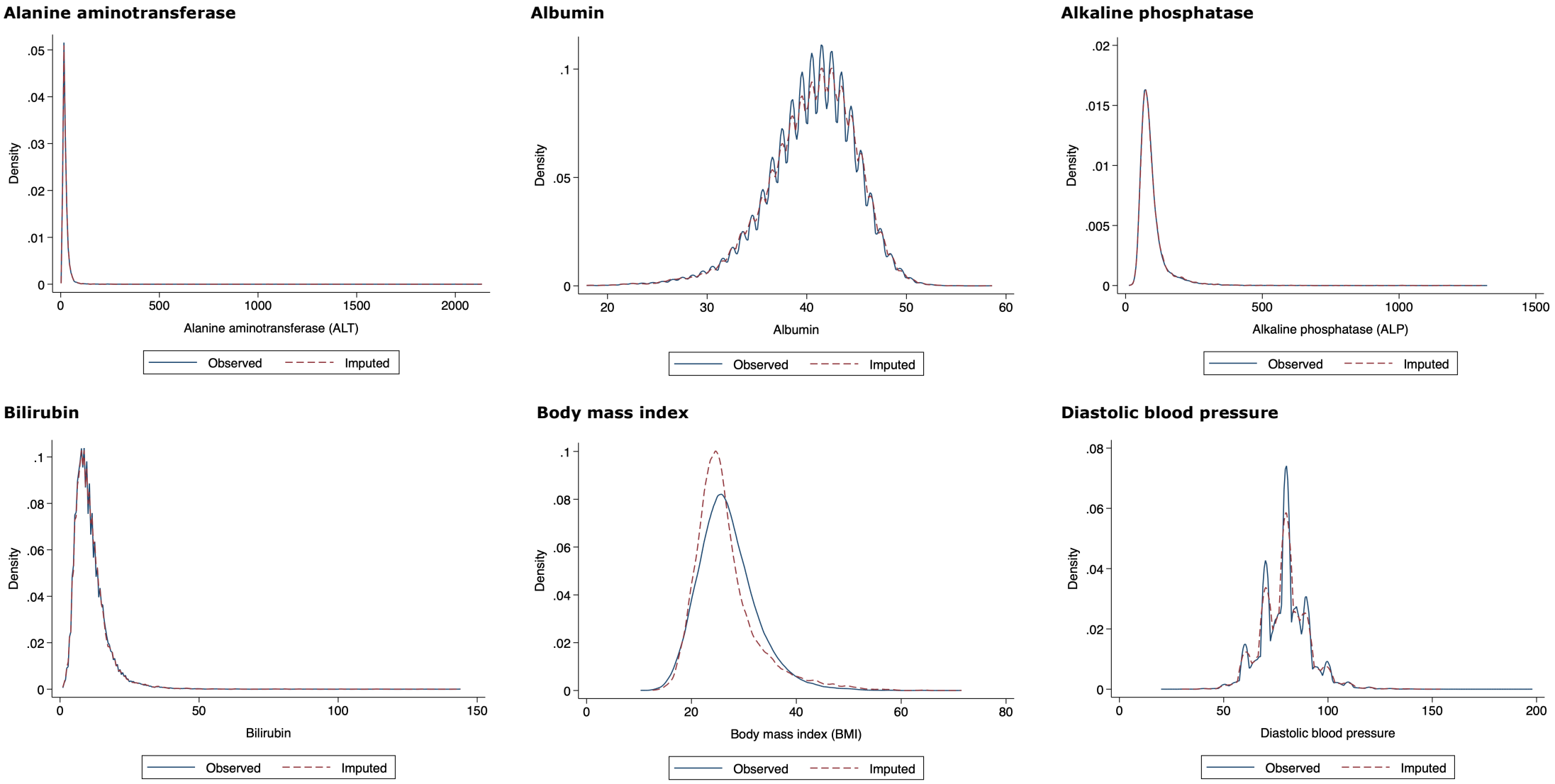


Figure 2.5a **Kernel density plots of observed versus imputed values for variables**

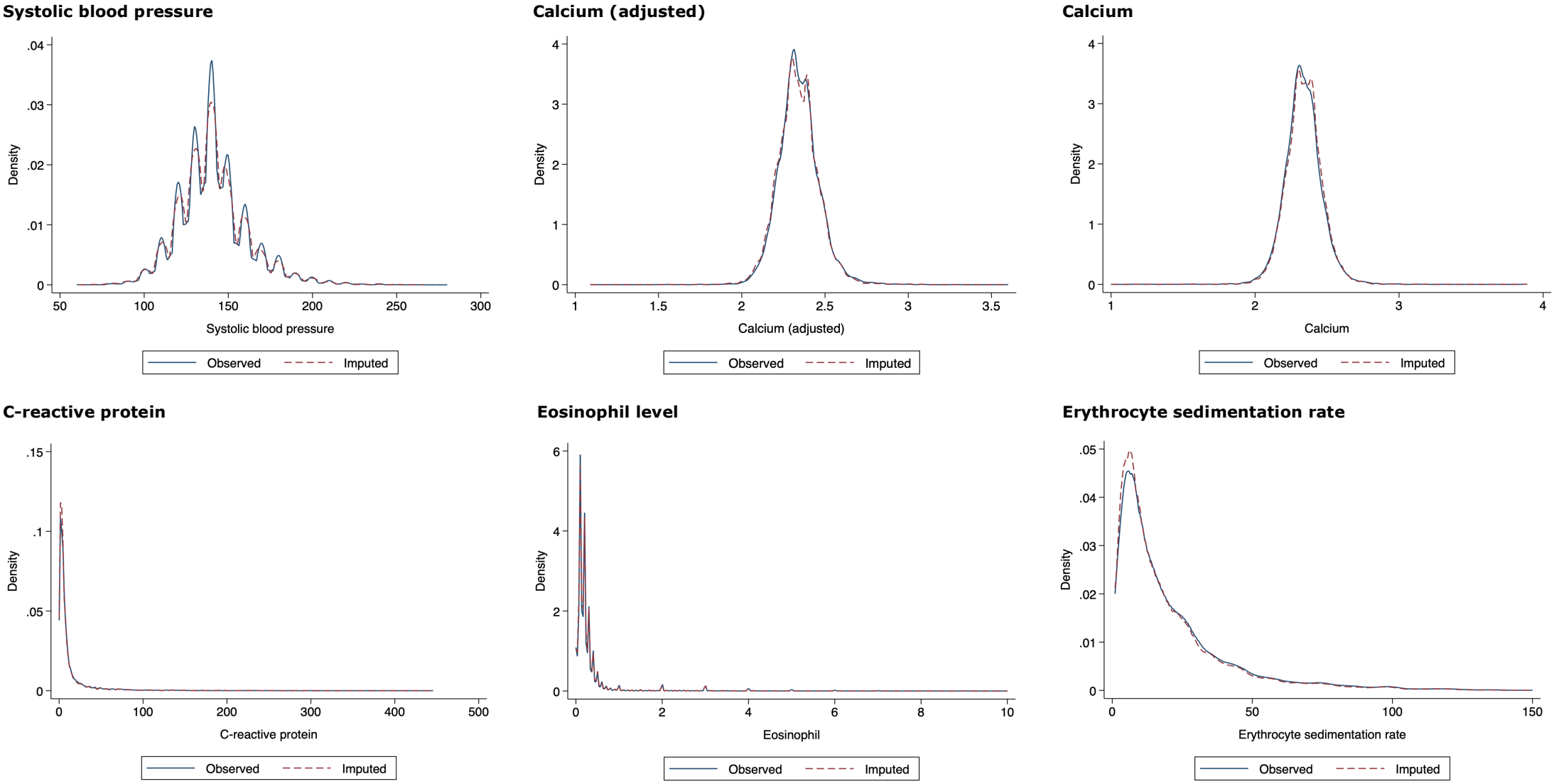
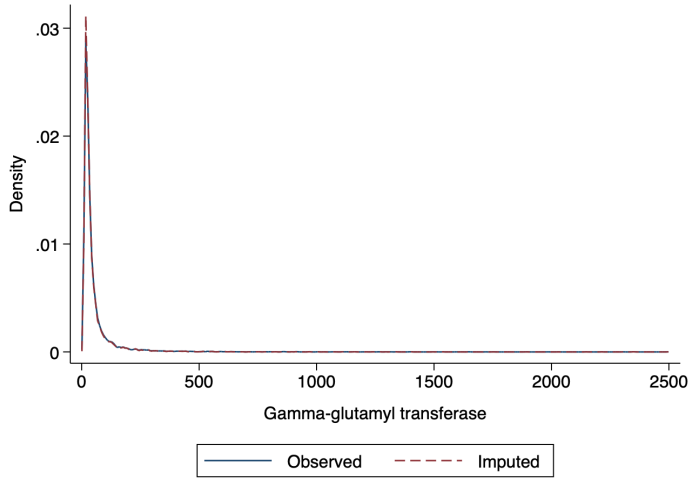
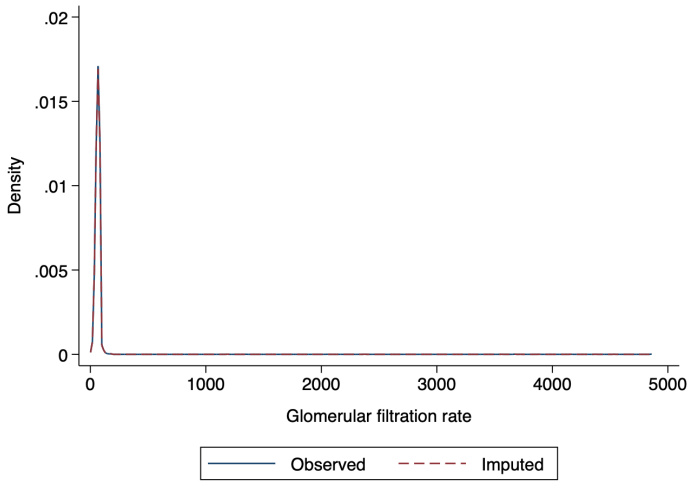


Figure 2.5b **Kernel density plots of observed versus imputed values for variables**

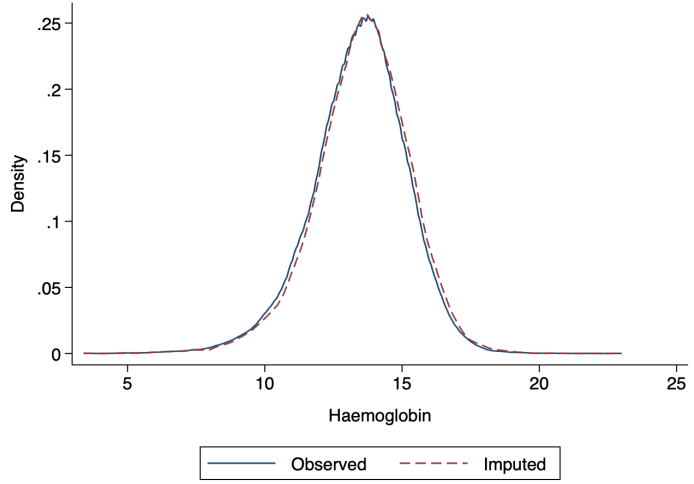
Gamma glutamyl transpeptidase



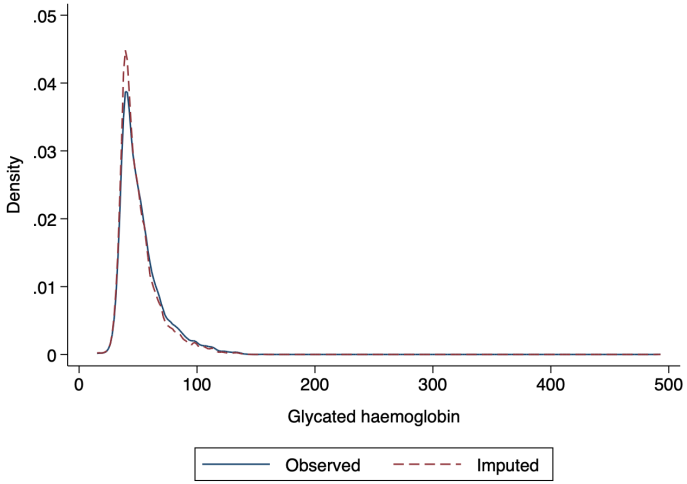
Glomerular filtration rate



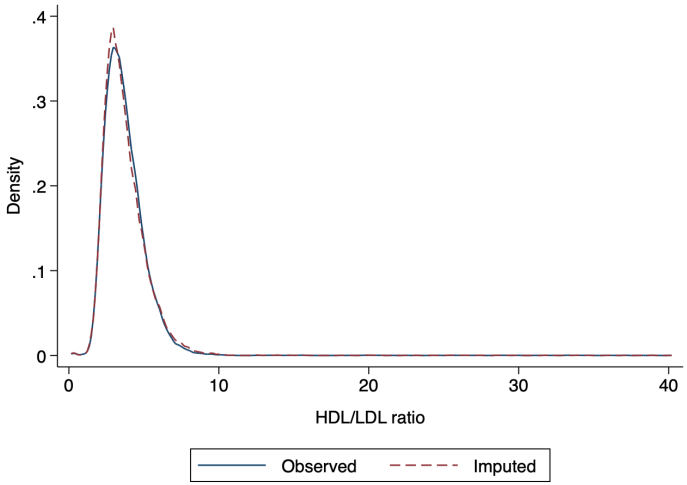
Haemoglobin level



Glycated haemoglobin



HDL/LDL ratio



Height

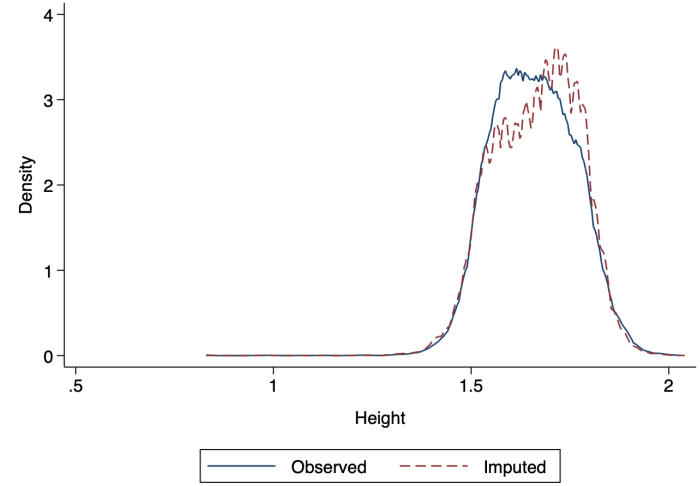
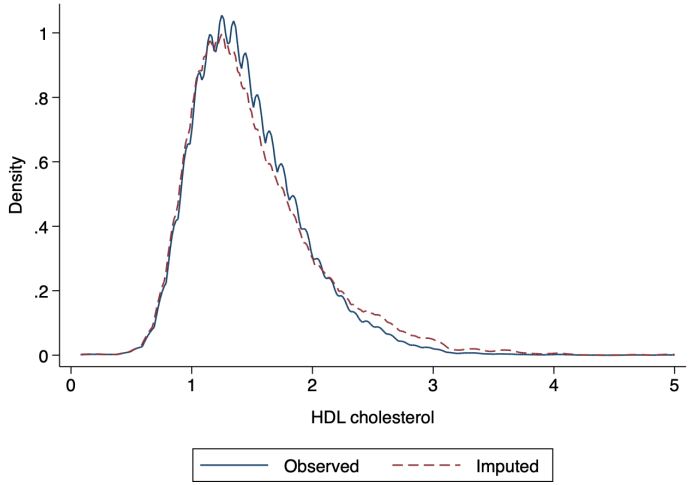
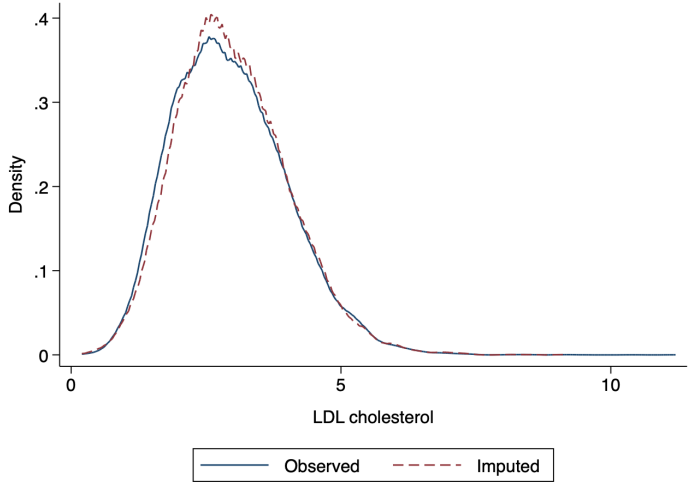


Figure 2.5c Kernel density plots of observed versus imputed values for variables

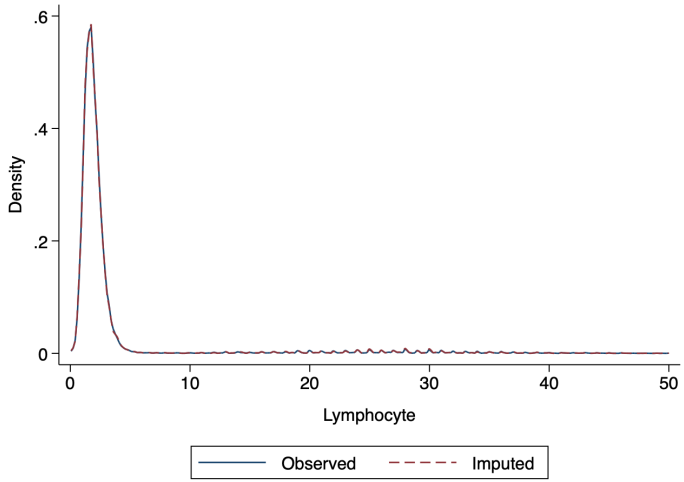
HDL cholesterol



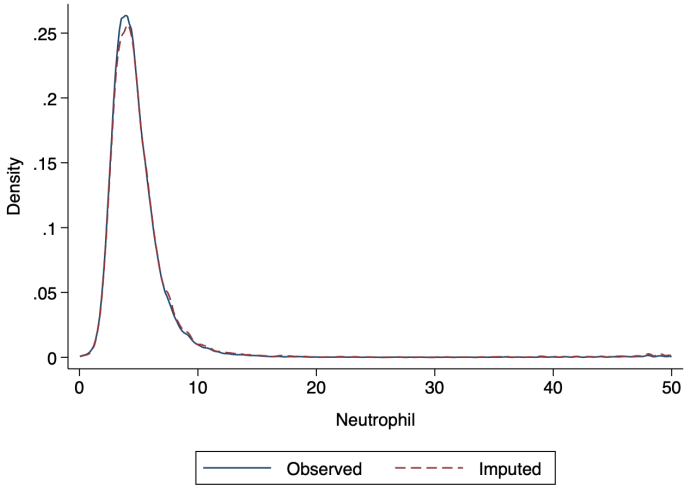
LDL cholesterol



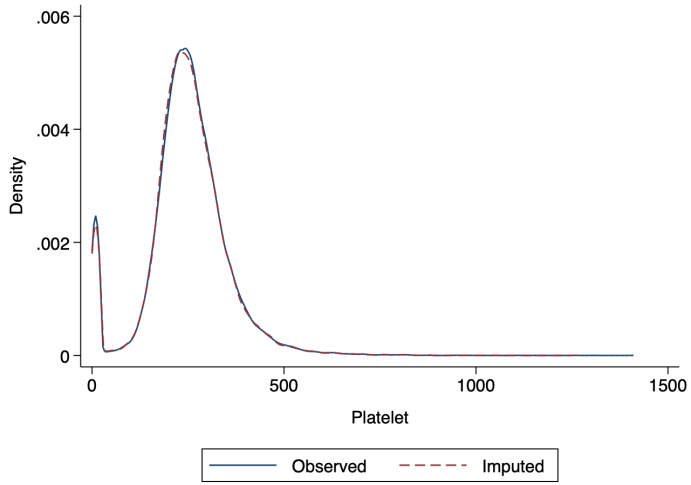
Lymphocyte count



Neutrophil count



Platelet count



Potassium level

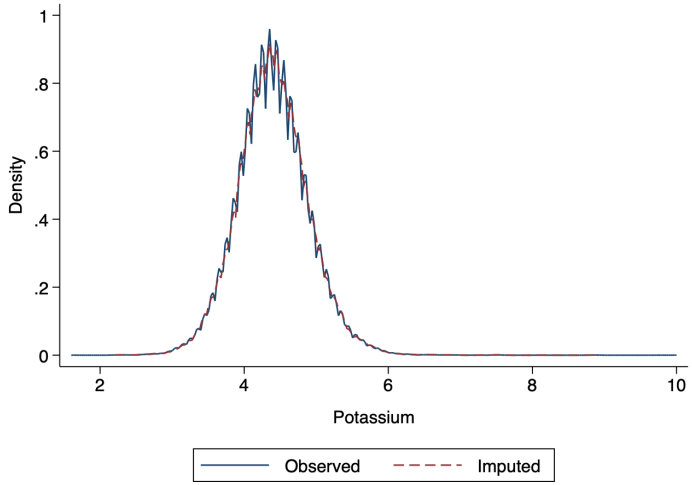


Figure 2.5d **Kernel density plots of observed versus imputed values for variables**

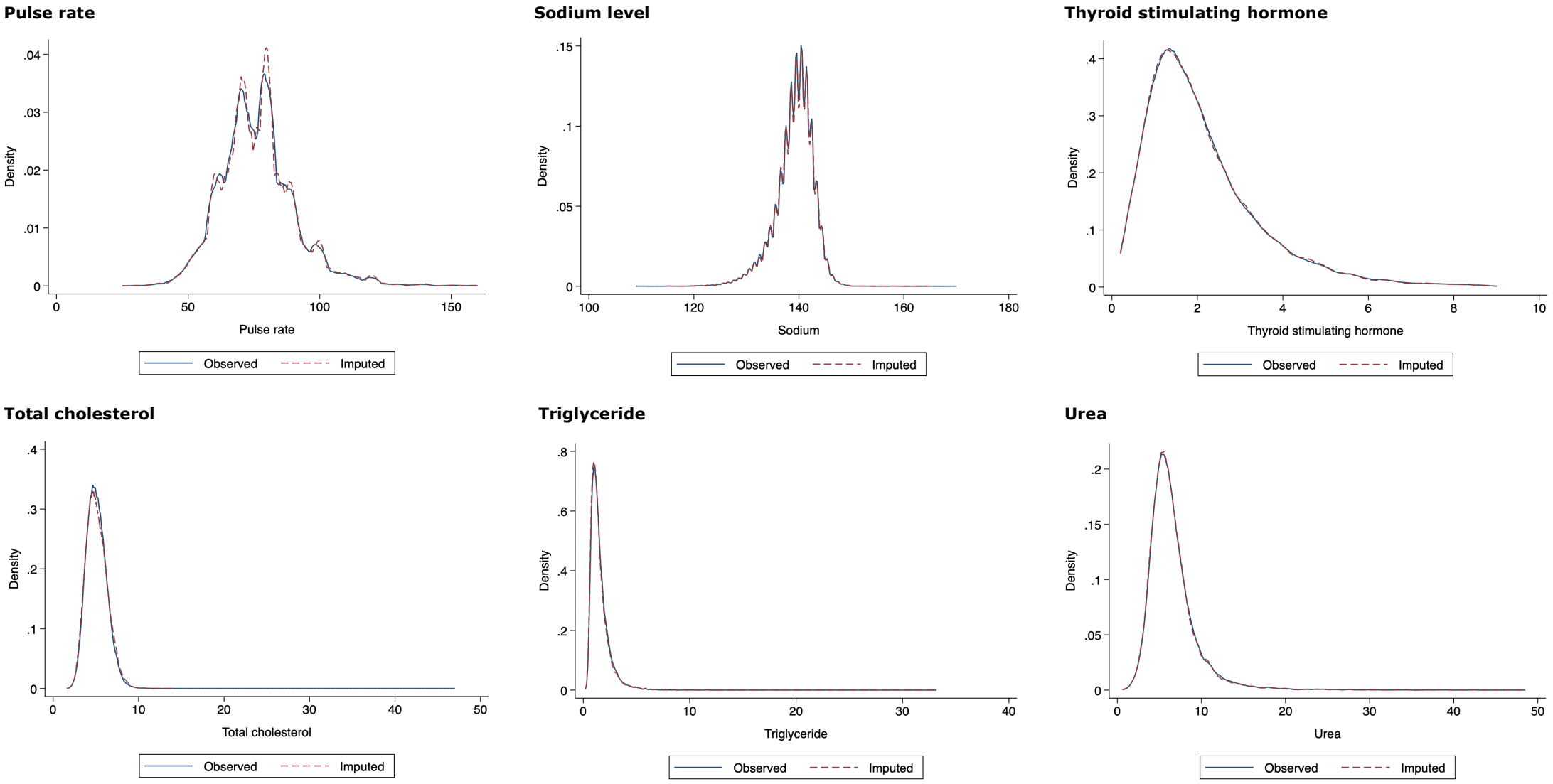


Figure 2.5e **Kernel density plots of observed versus imputed values for variables**

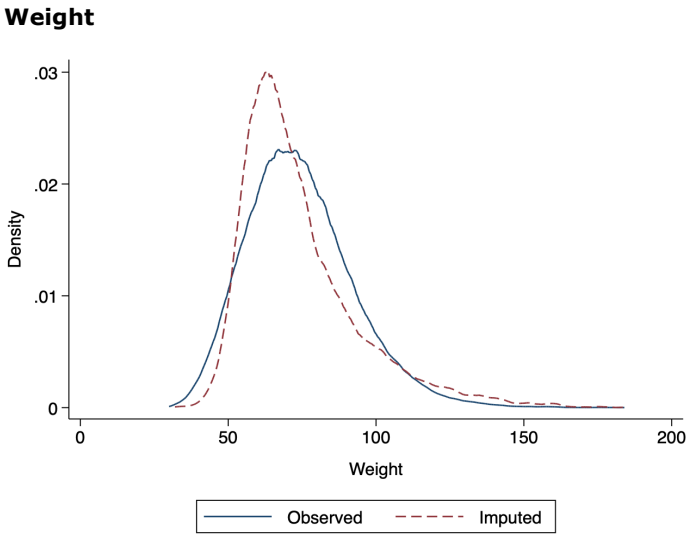


Figure 2.5f Kernel density plots of observed versus imputed values for variables

2.3.5 Limitations with using electronic health records

Despite the quality and size of these linked electronic health records, some limitations need to be highlighted:

- Relevant information that is useful for most cohort studies such as the use of over-the-counter medications, data on occupation, lifestyle habits, diet, and physical activities levels are generally not available, in being recorded sufficiently or in a consistently measurable/qualifiable way.
- Misclassification and hence the possibility of information bias cannot be ruled out. Exposures and risk factors such as smoking status and BMI may be infrequently recorded or only recorded when the patient was registering with the general practice.[112] The validity of most of the clinical diagnoses codes used for epidemiological research is yet to be assessed.[113,114]
- Due to issues related to confidentiality, potentially useful consultation information in the free-text form is not made available to researchers.
- For secondary care (hospital admission records), information on prescribed medications and diagnostic procedures are not always available.

Ethnicity recording has over the years been carried out in an ad-hoc manner, resulting in incomplete and unvalidated data. Hence, the quality of ethnicity data recording has been variable. Some primary care trusts have invested in the collection of ethnicity data to improve the quality sf primary care EHR.[115] Other innovative ways of improving the routinely collected data are being developed using new functionalities within GP systems.[116] This functionality enables clinicians to access a variety of communication methods for patients to provide needed information based on the patient's preference to means of communication – to receive SMS, Email, or via the patient app.

The Sentinel Stroke National Audit Programme, national stroke register, has data on about 90% of all stroke admissions in England and Wales. Linkage of primary care records to datasets such as the national stroke register could potentially provide more information on disease severity, treatment modalities, and other relevant information to compliments the information available in primary care databases.

2.4 Study designs and analyses

The methods used in this thesis research involves a number of statistical, epidemiological and data science techniques. A brief overview of the study methods used is provided in this section. Detailed information on methods is provided in the individual respective methods sections of subsequent chapters (chapters 3 – 7).

2.4.1 Systematic review and meta-analysis

Systematic reviews and meta-analyses use pre-planned explicit and reproducible methods to systematically search, critically appraise, and synthesise data from different studies conducted on a specific research topic.[117] Meta-analysis uses statistical methods to analyse, combine and summarize the results of the primary

studies.[118] In addition to providing up-to-date evidence-based information used in developing clinical guidelines, systematic reviews provide justification and inform research studies. Systematic review methodology was used in Chapter 3 to summarise the available evidence on prognostic models and assess their accuracy in predicting MACE outcomes in adults with an established stroke diagnosis.

2.4.2 Incidence rate estimation

Incidence rate, an epidemiological measure of disease frequency, is fundamental to monitoring disease conditions, formulating and evaluating healthcare interventions/policies.[119] The incidence of disease conditions in the general population also serves as an important indicator of a population's health status.[120] The comparison of incidence rates between studies and countries, and determining factors that explain these differences, results in increased knowledge on both prevention and aetiology of diseases.[121] Incidence is a rate of occurrence and thus related to a longitudinal design.[122] In Chapter 4 the incidence of non-fatal stroke and subsequent MACE outcomes were calculated. The incidence rates were subsequently stratified by age, sex, and socioeconomic status.

2.4.3 Propensity-score matching

Propensity score allows observational (non-randomised) studies to be designed and analysed with some peculiar characteristics of a randomised controlled trial (RCT). In RCTs, the gold standard approach for estimating the effects of treatments (exposures), random exposure allocation ensures that exposure status is not confounded with either measured or unmeasured baseline characteristics[123]. Propensity score serves as a balancing score that ensures similar distribution of observed baseline covariates/characteristics between the exposed and unexposed patients. There are 4 different propensity score methods[124]:

- *Propensity score matching*

Entails forming matched sets of exposed and unexposed individuals who share a similar value of the propensity score.[125,126]

- *Stratification on the propensity score*

  Involves stratifying individuals into mutually exclusive subsets based on their estimated propensity score.

- *Inverse probability of treatment weighting using the propensity score*

  The propensity score uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of exposure assignment.

- *Covariate adjustment using the propensity score*

  The outcome variable is regressed on an indicator variable denoting exposure status and the estimated propensity score.

Propensity-score matching was used in Chapter 5 to ensure similar distribution of observed baseline characteristics between the patients with incident haemorrhagic and ischaemic stroke. The risk of subsequent cardiovascular morbidity and mortality outcomes between the two patient groups was then assessed.

2.4.4 Landmark analysis

In the analysis of time-to-event data, the landmark method refers to the practice of designating a point occurring during the follow-up period (i.e., the landmark time) and analysing only those individuals who have survived until the landmark time.[127,128] Landmark analysis is used to avoid immortal time bias – occurring when a time-dependent exposure is not appropriately included in the analysis of time-to-event (survival) outcomes.[129] Landmark analysis was used in Chapter 5 to minimise the impact of incident stroke severity (a variable not available in the linked dataset) on subsequent outcomes.

2.4.5 Cox proportional hazards regression

The Cox (proportional hazards) regression model[130] is a commonly used approach for analysing survival (time-to-event) data in medical research.[131] Cox model is a survival analysis regression model, describing the relation between the event incidence, expressed by the hazard function and a set of covariates. The hazard function is the probability that an individual under observation experiences the outcome of interest in a period centred around that point in time.[132] Multivariable regression models estimate the relationship between a dependent variable (i.e., an outcome) and more than 1 independent variable. The most preferable and optimal way for covariate (confounder) selection is to specify in advance the dependent variables that will be included in the model based on expert clinical evidence.[133] Multivariable Cox regression was used in:

- Chapter 5: To estimate the association between incident stroke sub-types (haemorrhagic and ischaemic) and subsequent MACE outcomes.
- Chapter 6: To estimate the association between BMI and subsequent MACE outcomes.
- Chapter 7: To evaluate the association between the phenotypic clusters and subsequent cardiovascular morbidity and mortality outcomes.

To counteract the possible residual confounding effect on the findings, specific sub-group analyses were done. Most sub-group analyses were done to assess the robustness of study findings as well as confirm the findings.

2.4.6 Cluster analysis

Cluster analysis (clustering) is a method for detecting patterns and structures in both labelled and unlabelled datasets. Clustering has been used in many contexts and disciplines, including medicine, and shown to be useful in discovering unique groups/populations. In supervised machine learning, a set of input variables are mapped to a target outcome using a function in the process of approximating. The

term "supervised" refers to the process of the algorithm being supervised by having the target outcome (i.e., correct answers). However, for "unsupervised" only the set of input variables are available with no corresponding output variable (i.e., the data being unlabelled). Hence, for unsupervised machine learning, the algorithm discovers the structures in the datasets without the target outcome(s).[134] Clustering is one of the most important unsupervised learning techniques. Clustering aims to identify subgroups within heterogeneous data such that each cluster has greater homogeneity than the whole.[135] Clustering algorithms could find patterns across patients that may be difficult for medical practitioners to find.

The cluster analysis process can be divided into the following steps[134]:

1. *Pattern representation*

   This involves the pre-processing of extraction and selection of variables (features) to be used in the cluster analysis. The type, weight, and scale of the features for the clustering algorithm are defined at this stage. The use of weights to translate the importance of each feature leads to better clustering.[136] Feature scaling applies a mathematical transformation to each feature, and this ensures that all features make a comparable contribution to the measurement of similarity.

2. *Similarity measure*

   For two data points, the similarity measure quantifies how similar these points are and hence provides an indication of proximity, likeness, affinity, or association. Based on the variety of feature types and scales, there are several similarity measures available. However, there are two main types: metric and probability distribution-based similarity measures. Euclidean distance metric, the most popular metric used, is a measure of the geometric distance between two data points.

3. *Clustering algorithm*

The choice of clustering algorithm influences the clustering results for the data under analysis and the computational speed. The clustering algorithms can be broadly classified as:

- Hierarchical clustering: transforms a distance matrix of pairwise similarity measurements between all items into a hierarchy of nested groupings. The hierarchy is represented with a binary tree-like dendrogram that shows the nested grouping of patterns and the similarity levels at which groupings change.

- Partitional clustering: partitional clustering algorithm such as k-means (simplest and most commonly used algorithm) is a single partition of the data into a set of disjoint clusters. Partitional methods are used when the analysis involves very large data sets for which the construction of a dendrogram is computationally prohibitive. A drawback of partitional algorithms is that the number of clusters must be specified. In partitional techniques, the clusters produced optimize a criterion function defined either locally, that is, on a subset of patterns, or globally, so defined over all of the patterns.

- Density-based clustering: groups neighbouring objects of a data set into clusters based on density conditions measured in terms of the local distribution of nearest neighbours. Density-based algorithms typically assign clusters in dense regions of objects in the data space that are separated by regions of low density. Density-based algorithms are capable of discovering clusters of arbitrary shapes, providing natural protection against outliers. Some examples include Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[137] and DENsity CLUstEring (DENCLUE).[138]

- Grid-based clustering: mainly proposed for spatial data mining and it inherits the topology from the underlying attribute space. These algorithms divide the spatial area into a finite number of rectangular cells, generating several levels of cells corresponding to different levels of resolution, and then perform all operations on the quantized spatial area, which has the advantage of limiting the search combinations.

There are a number of subtypes and algorithms for each of the aforementioned cluster categories. Other classification criteria for clustering exists such as hard or fuzzy (soft); probabilistic or deterministic.

4. Assessment of the output

   The clusters defined are unknown *a priori*, hence need to be evaluated. There are two ways of validating the identified clusters[139]:

   - *External validation*: compares the identified clusters from the cluster analysis to a reference result considered to be the ground truth. If the identified clusters are somehow similar to the reference, the final output is considered a "good" clustering. This validation is straightforward when the similarity between the two has been well-defined. However, in most real-world applications, the reference result is mostly not available. External evaluation is, therefore, largely used for synthetic data and mostly for tuning clustering algorithms.

   - *Internal validation*: the evaluation of the cluster analysis is compared only with the result itself, i.e., the structure of found clusters and their relations to each other. This is a more realistic and efficient validation approach in many real-world applications as it does not refer to any assumed references from outside which is

not always feasible to obtain in most real-world applications. With the significant increase of the size and dimensionality of available data, complete knowledge of the ground truth is either unavailable or not always valid.

5. *Graphical representation*

   For easy interpretation, the cluster results are represented in graphical displays.

In Chapter 7, an unsupervised machine learning technique was used to explore heterogeneity in clinical characteristics of adult patients with incident stroke and cluster patients based on phenotypic similarities.

2.5 Statistical software used

All statistical analyses were performed using Stata SE version 16.1 or 17.0 (StataCorp LP) and R version 4.1.0 (http://cran.r-project.org). An alpha level of 0.05 was used for all analyses.

2.6 Power calculation

A key advantage of using the linked datasets (CPRD GOLD, HES APC, ONS mortality, and social deprivation data) is the very large size of the datasets. This enables precise estimation of effect sizes and the exploration of a wide range of potential risk factors for stroke and subsequent cardiovascular morbidity and mortality outcomes. Due to the large sample size for the respective studies, formal power calculations were not needed for the analyses.

Summary

This chapter described the data sources used in addressing the research questions in this thesis, the data management principles applied and provided an overview of the statistical analysis methods used. The chapters that follow present original studies that contribute to stroke research in the primary care setting. These studies provide evidence to better target care for patients with incident stroke.

Chapter 3

Prognostic prediction models for major adverse cardiovascular events in adults with stroke: A systematic review

This chapter presents the findings of a systematic review that explored prognostic models for predicting MACE outcomes in patients with an established diagnosis of stroke. A broader search strategy on patients with an established diagnosis of cardiovascular disease (defined as either CHD, stroke, or PVD) was done. For this chapter, only results focused on patients with an established diagnosis of stroke is, however, discussed.

The protocol for this research study has been published in the journal *BMJ Open*:

Akyea RK, Leonardi-Bee J, Asselbergs FW, Patel RS, Durrington P, Wierzbicki AS, Ibiwoye OH, Kai J; Qureshi N, Weng SF. 2020. Predicting major adverse cardiovascular events for secondary prevention: protocol for a systematic review and meta-analysis of risk prediction models. BMJ Open, 10(7), https://doi.org/10.1136/bmjopen-2019-034564

3.1 Abstract

Background: Clinical guidelines recommend the stratification of patients with an established diagnosis of a stroke to guide subsequent management decisions. There is, however, no published evidence of the predictive value of existing risk prediction models, limiting the confidence with which these models could be used. This study aims to summarise the available evidence on prognostic models and assess their accuracy for predicting MACE outcomes in adults with established diagnosis of stroke.

Methods: MEDLINE, EMBASE, PsycINFO, and Web of Science were searched from inception to April 2020. Peer-reviewed studies developing, validating, or updating a multivariable prognostic model for subsequent MACE outcomes in adults ≥16 years with an established diagnosis of stroke were identified. Two reviewers independently screened, extracted relevant data, and assessed the risk of bias using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). A narrative synthesis was conducted. Subsequent MACE outcome was defined as a composite of either CHD, stroke, PVD, heart failure, or CVD-related mortality.

Results: Forty eligible articles with 23 distinct prognostic models were identified – describing the development of 11 prognostic models and 75 external validations. Among the 23 models, the most frequently used predictors were age, sex, history of transient ischaemic attack, hypertension (blood pressure), and diabetes. There were methodological limitations in the development of the prognostic models – improper or no information on the handling of missing data, selection of candidate predictor variables, or incomplete evaluation of model performance (no model calibration). All the development models had a high risk of bias. Model predictive accuracy, measured by the area under the receiver operating curves or c-statistic, ranged from 0.632 (95% CI: 0.579 – 0.684) for the Modified Essen Stroke Risk model predicting recurrent ischaemic stroke within a year of stroke event to 0.85

(95% CI: 0.78 – 0.91) for the ABCD score predicting recurrent stroke within 7 days of stroke event for the model development studies. For the validation studies, the Recurrence Risk Estimator model had the best predictive accuracy of 0.86 (95% CI: 0.80 – 0.93) for the prediction of recurrent ischaemic stroke within 90 days of the initial stroke event.

Conclusions: Many prognostic models have been developed and validated for predicting subsequent cardiovascular morbidity and mortality outcome in patients with stroke. The clinical utility of these prognostic models, however, remains uncertain due to methodological limitations.

Systematic review registration: PROSPERO CRD42019149111

3.2 Introduction

The prevention of CVD (including stroke), the leading cause of morbidity and mortality,[140] represents one of the most important aspects of preventive medicine. More people are surviving initial CVD events[141,142] and for patients with established CVD, the priority is to prevent a subsequent CVD event or premature death. Current secondary prevention interventions have achieved substantial success in reducing the risk of cardiovascular events and mortality after incident stroke.[143] However, the prognosis of patients with established stroke remains sub-optimal.[144,145]

Prognostic prediction models are generally equations, based on routinely collected clinical information, that converts a combination of predictor variables to an estimate of an individual's risk of developing a defined outcome over a specified period.[146] Prognostic prediction models might be an effective tool for risk stratification. Although identifying patients at risk could facilitate secondary preventive strategies, guide therapy, and help in clinical research, there is no

previous review of the literature exploring the validity of risk prediction models for MACE outcomes after incident stroke events. This review aimed to, therefore, provide an overview of the prognostic risk models developed, validated, or updated, their composition, and discriminatory accuracy for predicting MACE outcomes in adults with an established diagnosis of stroke.

3.3 Methods

This systematic review was conducted according to the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS)[147] and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.[148] A protocol for this review has previously been published[149] and registered on PROSPERO (registration number CRD42019149111). Ethical approval and patient informed consent was not needed as all data were obtained from previously published studies.

Literature search

Bibliographic databases (Ovid MEDLINE, EMBASE, PsycINFO, and Web of Science) were searched from database inception to April 2020. The employed search terms are presented in Appendix C.3.1 and aimed to cover expressions of cardiovascular disease, risk scores, and discriminatory accuracy. The references of each eligible article and its citations on Google Scholar were also searched to identify potential papers that fulfilled the eligibility criteria.

Eligibility criteria

All studies reporting the development and validation of at least one multivariable model for predicting the risk of MACE outcome in adult patients with an established diagnosis of CVD were included. Box 3.1 provides a detailed description of the PICOTS for this review.

Box 3.1 Key items for framing aim, search strategy, and study inclusion and exclusion criteria for systematic review – following PICOTS guidance

PICOTS	Inclusion and Exclusion Criteria
Populations	Adults, 16 years and above, with an established diagnosis of cardiovascular disease (CVD) [a]
Interventions	Any multivariable prognostic model for making individualised predictions.
Comparators	Not applicable
Outcomes	Major adverse cardiovascular event (MACE) [b] Included studies should report results for at least one of the components of MACE.
Timing	Predictors measured at any timepoint in clinical course of CVD and preceding outcome; any duration of follow-up for outcome without applying any specific limitation in prediction horizon [c]
Setting	Any setting – in-patient, out-patient, and community
Study design	Comparative study designs including clinical trials, cohort, case-control, and cross-sectional studies.

[a] CVD was defined as a documented clinical diagnosis of arterial occlusive events including coronary artery disease, cerebrovascular artery disease and peripheral vascular disease (PVD).

[b] Major adverse cardiovascular event defined as a diagnosis of either coronary artery disease (including myocardial infarction, coronary artery bypass grafting (CABG), percutaneous coronary intervention (PCI); stroke (including carotid endarterectomy); peripheral vascular disease (including PVD-related complications such as gangrene, amputation); heart failure; or CVD-related mortality.

[c] Follow-up was categorised as: short-term (≤ 1 year); medium-term (1–5 years); and long-term

The eligible studies either reported the development of the multivariable model(s), external validation of an existing model(s), and/or the update of existing model(s).[150] Further included studies explicitly estimated and presented a measure of the model's performance. Eligible studies reported original research, human studies, and there was no language restriction.

Studies that developed models to identify patients with existing MACE outcomes were excluded[6]. Given that prognostic models estimate the probability of an outcome for an individual patient over a period, cross-sectional studies were excluded because predictors and the outcome are all measured at the same time. Cohort studies that did an external validation of a model derived from a cross-sectional study were, however, considered eligible.

Data extraction

To facilitate the data extraction process, a standardised form based on the CHARMS checklist[147] recommendations was constructed. When a study described a model's performance both in an overall sample and in specific subgroups, the analysis of the total population was extracted. For articles describing multiple models, data was extracted separately for each model.

Risk of bias assessment

The risk of bias and applicability concerns of the included studies, either developing or externally validating prognostic models, were assessed using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) – a risk of bias assessment tool designed for systematic reviews of diagnostic or prognostic prediction models.[151]

Two independent reviewers screened all articles (title and abstract as well as full text), extracted data from the articles and assessed the risk of bias for included articles. Any initial disagreement was resolved through further discussion among the reviewers. I cross-checked all the extracted data and risk of bias assessment.

[6] These studies (diagnostic prediction models) estimate the probability that a disease condition is currently present or absent for an individual. Prognostic models, however, estimate whether an individual will experience a specific outcome or condition in the future (a specified time).

Statistical analysis

The descriptive statistics are reported to summarise the characteristics of the models. The discriminatory accuracy of a prognostic risk model (i.e., its ability to distinguish between patients developing and not developing MACE outcomes of interest) was measured by the area under the receiver operating characteristic curve (AUC) or C-statistic, which ranges from 0.5 (no discriminative ability) to 1.0 (perfect discriminative ability).

3.4 Results

The initial electronic search generated 4,178 records, including 1,615 citations from OVID Medline, 1,217 citations from OVID EMBASE, 75 citations from OVID PsycINFO and 1,271 citations from Web of Science. After removal of duplicates (n=1,499), the titles and abstracts of 2,679 citations were screened using Covidence – a web-based systematic review management tool. Screening of titles and abstracts resulted in 218 relevant articles. After the full-text screening, 105 articles met eligible criteria. Detailed information on the selection process is presented in the PRISMA flow diagram, Figure 3.1.

3.4.1 Prognostic prediction models for stroke

This review identified 40 peer-reviewed studies that either developed[7] or validated[8] prognostic prediction models for MACE outcomes in patients with

[7] Development: Creating a risk prediction model using patient characteristics to estimate the probability that a certain outcome is present (diagnostic) or will occur within a defined time (prognostic model). This includes identifying and selecting classifying variables and assessing model performance.

[8] Validation: External validation is when an already developed risk prediction model is applied to an independent sample from a comparable population to estimate the risk of the outcome, compared observed outcome, and the performance of the model.

incident stroke. Of the 40 peer-reviewed studies included, 7 performed model development only and 4 performed model development with validation. In total there were 75 prognostic model validations reported.

Twenty-three (23) distinct multivariable prognostic models were, however, extracted from the 40 studies. Table 3.1 provides the details for the prognostic models assessed in the included studies.

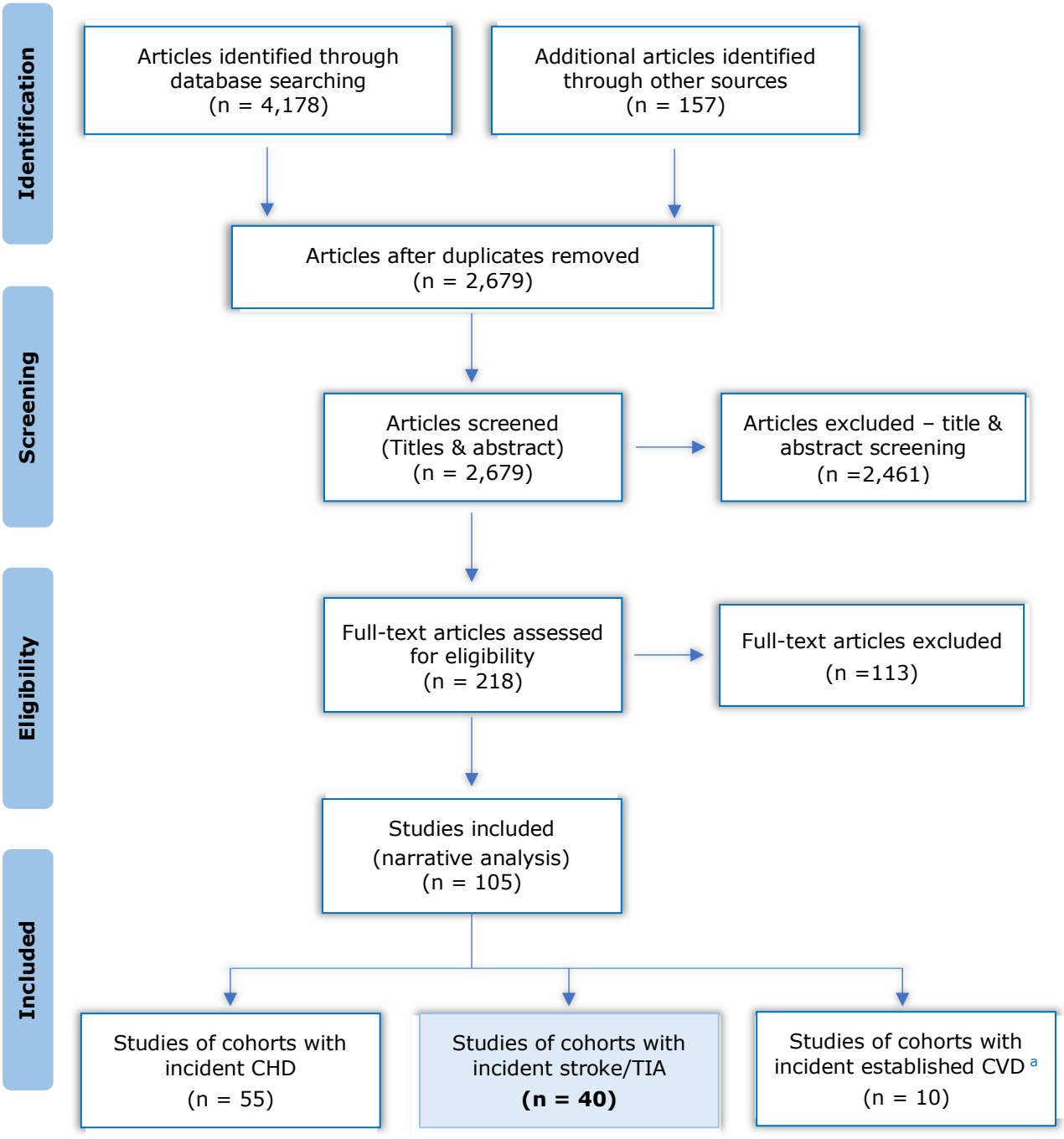


Figure 3.1 PRISMA Chart (flow diagram of the selection process)

Number of articles identified from the various databases searched – Medline: 1,615; EMBASE: 1,217; PsycINFO: 75; Web of Science: 1,271.

[a] Established CVD was defined in these studies as a clinical manifestation of arterial disease – CHD disease, cerebrovascular disease, PVD, or abdominal aortic aneurysm.[152]

Table 3.1 Prognostic models for major adverse cardiovascular event prediction in patients with established stroke

Reference	Models assessed	Development vs. Validation
Andersen 2015 [153]	CHA2DS2VASc score	Validation
Andersen 2015 [153]	Essen Stroke Risk Score	Validation
Andersen 2017 [154]	CHA2DS2VASc score	Validation
Andersen 2017 [154]	Essen Stroke Risk Score	Validation
Andersen 2017 [154]	CHA2DS2VASc score + DWMH score ≥2	Validation
Andersen 2017 [154]	Essen Stroke Risk Score + DWMH score ≥2	Validation
Andersen 2017 [154]	DWMH score	Validation
Andersen 2017 [154]	PVH score	Validation
Andersen 2017 [154]	Total Fazekas score	Validation
Arsava 2011 [155]	Recurrence Risk Estimator score	Validation
Arsava 2011 [155]	$ABCD_2$ score	Validation
Arsava 2016 [156]	Recurrence Risk Estimator score	Validation
Asimos 2010 [157]	$ABCD_2$ score	Validation
Ay 2010 [158]	Recurrence Risk Estimator score	Development and validation
Ay 2010 [158]	Stroke Prognosis Instrument II	Validation
Ay 2010 [158]	Essen Stroke Risk Score	Validation
Bhaskar 2017 [159]	Model 1	Development
Bray 2007 [160]	ABCD score	Validation
Chandratheva 2010 [161]	$ABCD_2$ score	Validation
Chandratheva 2011 [162]	$ABCD_2$ score	Validation
Chandratheva 2011 [162]	Essen Stroke Risk Score	Validation
Chandratheva 2011 [162]	Stroke Prognosis Instrument II	Validation
Chatzikonstantinou 2013 [163]	$ABCD_2$ score	Validation
Chatzikonstantinou 2013 [163]	$ABCD_3I$ score	Validation
Chen 2016 [164]	Essen Stroke Risk Score	Validation
Coutts 2008 [165]	$ABCD_2$ score	Validation
Coutts 2008 [165]	MRI	Validation
Coutts 2008 [165]	$ABCD_2$ score +MRI	Validation
Coutts 2008 [165]	All factors	Validation
Fothergill 2009 [166]	ABCD score	Validation
Fothergill 2009 [166]	$ABCD_2$ score	Validation
Ghia 2012 [167]	$ABCD_2$ score	Validation
Hakan 2009 [168]	$ABCD_2$ score	Validation
Hakan 2009 [168]	DWI information	Validation
Hakan 2009 [168]	Dichotomized $ABCD_2$ score with DWI	Development
Hakan 2009 [168]	CIP model	Validation
Johnston 2007 [169]	California Score	Validation
Johnston 2007 [169]	ABCD score	Validation
Johnston 2007 [169]	$ABCD_2$ score	Validation
Kamouchi 2012 [170]	Fukuoka stroke risk score	Development
Kernan 2000 [171]	Stroke Prognosis Instrument I	Validation

Reference	Models assessed	Development vs. Validation
Kernan 2000 [171]	Stroke Prognosis Instrument II	Development and validation
Ling 2018 [172]	Modified Essen Stroke Risk Score	Development
Ling 2018 [172]	Essen Stroke Risk Score	Validation
Liu 2013 [173]	Essen Stroke Risk Score	Validation
Liu 2013 [173]	Stroke Prognosis Instrument II	Validation
Liu 2017 [174]	Essen Stroke Risk Score	Validation
Maier 2013 [175]	Essen Stroke Risk Score	Validation
Maier 2013 [175]	$ABCD_2$ score	Validation
Maier 2013 [175]	Recurrence Risk Estimator score	Validation
Meng 2011 [176]	Essen Stroke Risk Score	Validation
Meng 2011 [176]	Stroke Prognosis Instrument II	Validation
Purroy 2012 [177]	ABCD score	Validation
Purroy 2012 [177]	$ABCD_2$ score	Validation
Purroy 2012 [177]	$ABCD_2I$ score	Validation
Purroy 2012 [177]	ABCDI score	Validation
Purroy 2012 [177]	$ABCD_3$ score	Validation
Purroy 2012 [177]	$ABCD_3V$ score	Validation
Purroy 2012 [177]	Essen Stroke Risk Score	Validation
Purroy 2012 [177]	Stroke Prognosis Instrument II	Validation
Purroy 2012 [177]	California Risk Score	Validation
Rothwell 2005 [178]	ABCD score	Development and validation
Sanders 2011 [179]	$ABCD_2$ score	Validation
Sciolla 2008 [180]	ABCD score	Validation
Sciolla 2008 [180]	ABCDI score	Validation
Sheehan 2009 [181]	$ABCD_2$ score	Validation
Sheehan 2010 [182]	$ABCD_2$ score	Validation
Song 2013 [183]	$ABCD_3I$ score	Validation
Song 2013 [183]	$ABCD_2$ score	Validation
Song 2015 [184]	$ABCD_2$ score	Validation
Song 2015 [184]	Recurrence Risk Estimator score	Validation
Sumi 2013 [185]	Modified Essen Stroke Risk Score	Development and validation
Tsivgoulis 2010 [186]	$ABCD_2$ score	Validation
Weimar 2008 [187]	Essen Stroke Risk Score	Validation
Weimar 2008 [187]	Ankle Brachial Index	Validation
Weimar 2009 [188]	Essen Stroke Risk Score	Validation
Weimar 2010 [189]	Essen Stroke Risk Score	Validation
Weimar 2010 [189]	Hankey score	Validation
Weimar 2010 [189]	LiLAC score	Validation
Weimar 2010 [189]	Stroke Prognosis Instrument II	Validation
Weimar 2012 [190]	Essen Stroke Risk Score	Validation
Weimar 2012 [190]	Stroke Prognosis Instrument II	Validation
Wijk 2005 [191]	Model 1	Development
Wijk 2005 [191]	Model 2	Development
Wijk 2005 [191]	Model 3	Development
Yang 2010 [192]	$ABCD_2$ score	Validation

Model development

3.4.1.1 Characteristics of studies developing models

The characteristics of the 9 studies reporting 11 prognostic model developments are summarised in Table 3.2. Patient characteristics varied significantly across studies in terms of the study populations (that is, mean age and sex distribution), outcomes predicted, and follow-up duration/period for the various outcomes. Follow-up duration from stroke event to the outcome of interest for the studies ranged from 7 days [168,178] to 10.1 years in the study by Wijk *et al*., 2005.[191]

All the models were developed using cohort (prospective/retrospective) studies based on routine clinical data, except for one study (Kernan *et al*., 2000[171]) which used data from randomised controlled trials (RCTs). Most of the studies (7/9) were based in a hospital setting with only 2 studies including patients from a primary care setting. Data from 6 different countries were used in these models – Australia, China, Japan, Netherlands, the UK, and the USA. Missing data is a common problem with routine clinical data, only 1 model used the multiple imputation approach to deal with missing data,[168] 5 studies had no information on missing data and how it was dealt with,[158,159,171,178,191] and the remaining 3 studies did a complete-case analysis.[170,172,185] The study sample size used for model development ranged from 209 for the ABCD model[178] to 3,452 for the Modified Essen Stroke Risk Score model.[185]

For the selection of candidate predictors (risk factors), 5 studies used univariate analysis, 1 study used a combination of pre-specified risk factors based on clinical knowledge and univariate analysis, and the remaining did not specify what was done. Cox proportional hazards regression model was the most common statistical model used in the development of the prognostic models.

Table 3.2 Characteristics of studies developing prognostic models for major adverse cardiovascular events in patients with stroke

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study population	Outcome(s) predicted	Follow-up for outcome
Ay 2010 [158]	Recurrence Risk Estimator (RRE-90)	2003 - 2006	Retrospective	Clinical-based model: 1,485 Clinical- and imaging-based model: 1,257	Patients with ischaemic stroke • USA • Hospital setting • Age (years): – No recurrent stroke: (n=1,398): 72 (60-80) – Recurrent stroke (n=60): 74 (60-80) • Females: – No recurrent stroke: 649 (46%) – Recurrent stroke: 26 (43%)	Recurrent ischaemic stroke	90 days
Bhaskar 2017 [159]	Model 1	2006 - 2013	Retrospective	608	Patients with acute stroke: • Australia • Hospital setting • Age: 75.28 years (12.94) • Female: 292 (48%)	90-day stroke mortality	90 days
Hakan 2009 [168]	Dichotomized ABCD2 score with DWI	2000 - 2006	Retrospective	Recruited: 601 With follow-up data: 479	Patients with TIA: • USA • Hospital (ER) setting • Age: 67.7 years (14.7) • Female: 246 (51.6%)	Stroke	7 days
Kamouchi 2012 [170]	Fukuoka stroke risk score	2007 - 2011	Prospective and retrospective registry	Fukuoka Stroke Registry: 3,067	Patients with ischaemic stroke: • Japan • Hospital setting • Age: 71.8 years (12.2) • Males: 1,865 (60.8%)	Recurrent ischaemic stroke	1 year

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study population	Outcome(s) predicted	Follow-up for outcome
Kernan 2000 [171]	The Stroke Prognosis Instrument II	Not reported	RCTs	Women's Estrogen for Stroke Trial (WEST): 525 UK-TIA Aspirin Trial: 2449 CAPRIE Trial: 6431 Northern Manhattan Stroke Study (NoMaSS): 340 Total cohort: 9,678	Patients with ischaemic stroke & TIA: • UK, USA • Hospital setting • Age: Not reported • Female: Not reported	Composite outcome comprising stroke or death *(all-cause mortality)*	2 years
Ling 2018 [172]	Modified Essen Stroke Risk Score	2012 - 2014	Prospective	The First Affiliated Hospital of Jinzhou cohort: 773	Patients with acute ischaemic stroke: • China • Hospital setting • Age: 66 years • Female: 236 (20.5%)	Recurrent ischaemic stroke	1 year
Rothwell 2005 [178]	ABCD score	Development: Nov 1981 and Oct 1986 Validation: 2002-2004	Prospective	*Development:* Oxfordshire Community Stroke Project (OCSP): 209 *Validation:* Oxford Vascular (OXVASC) Study: 190	Patients with TIA: • UK (England) • Primary care setting • Age: – Development: 69.9 years (12.2) – Validation: 73.7 years (12.5) • Male: – Development: 112 years (54%) – Validation: 79 years (42%)	Stroke	7 days

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study population	Outcome(s) predicted	Follow-up for outcome
Sumi 2013 [185]	Modified Essen Stroke Risk Score	2007 - 2008	Prospective	EVEREST Prospective Ischemic Stroke Registry: 3,452	Patients with ischaemic stroke: • Japan • Primary care setting • Age (years): – No recurrent stroke *(n=3,171)*: 70 (62-76) – Recurrent stroke *(n=121)*: 71 (67-77) • Male: – No recurrent stroke: 2114 (67%) – Recurrent stroke: 89 (74%)	Recurrent ischemic stroke Cardiovascular event *(a composite of fatal/nonfatal stroke, MI, nonfatal unstable angina, and cardiac death)*	1 year
Wijk 2005 [191]	Model 1 Model 2 Model 3	DTT Aspirin Trial: 1986 - 89 then continued to 1990. All alive at end of study – followed to 2003	Prospective	LiLAC Cohort Study [*based on Dutch TIA Trial (DTT)]*: 2,473	Patients with TIA or minor stroke: • Netherlands • Hospital setting • Age: 65 years (10.1) • Male: 1489 (60.2%)	All-cause mortality Composite outcome comprising death from all vascular causes, non-fatal stroke, or non-fatal myocardial infarction	Mean: 10.1 years (SD 4.8)

The prognostic models were mainly presented as risk scores and risk equations. For the performance of the prognostic models, the discrimination (predictive accuracy) ranged from 0.632 (95% CI: 0.579 – 0.684) for the Modified Essen Stroke Risk predicting recurrent ischaemic stroke within a year of stroke event[185] to 0.85 (95% CI: 0.78 – 0.91) for the ABCD score predicting recurrent stroke within 7 days of stroke event[178] – Table 3.3. The majority (6/9) of studies did not provide information on model calibration.

3.4.1.2 External validations

There were 75 external validations in total for the 23 distinct multivariable prognostic models. The patient characteristics varied significantly across the validation studies – Appendix C.3.2. The study sample size for the validation studies ranged between 102 and 42,182 patients, mostly from hospital settings. The Recurrence Risk Estimator model had the best discrimination (predictive accuracy) of 0.86 (95% CI: 0.80 – 0.93) for the prediction of recurrent ischaemic stroke within 90 days of the initial stroke event. The discrimination for most of the models evaluated was poor – Appendix C.3.3. The ABCD2 score, Essen stroke risk score, and Stroke Prognosis Instrument II were the most validated prognostic models.

Table 3.3 Predictive accuracy of prognostic models developed for major adverse cardiovascular events in patients with stroke

Reference	Model	Cohort name	Sample size (cases/total)	Outcome	C-statistic (95% CI)	Calibration
Ay 2010 [158]	Recurrence Risk Estimator (RRE-90) *(Clinical-based model)*	728 patients with complete 90-day follow-up	60/1485	Recurrent ischaemic stroke	0.70 (0.63 - 0.77)	Calibration x^2 statistics p-value for lack of fit: 0.993
	RRE-90 *(Clinical-based model)*	433 patients from validation dataset	60/1485	Recurrent ischaemic stroke	0.70 (0.61 - 0.79)	Calibration x^2 statistics p-value for lack of fit: 0.993
	RRE-90 *(Clinical- and image-based model)*	728 patients with complete 90-day follow-up	54/1257	Recurrent ischaemic stroke	0.80 (0.73 - 0.86)	Calibration x^2 statistics p-value for lack of fit: 0.092
	RRE-90 *(Clinical- and image-based model)*	433 patients from validation dataset	54/1257	Recurrent ischaemic stroke	0.76 (0.66 - 0.87)	Calibration x^2 statistics p-value for lack of fit: 0.092
	RRE *(14-day risk, both datasets)*		30	Recurrent ischaemic stroke	0.80 (0.72-0.87)	No information
Bhaskar 2017 [159]	Model 1		126/608	Stroke mortality (90-day)	0.8 (-)	No information
Hakan 2009 [168]	DWI information		25/479	Stroke	0.76 (0.67-0.86)	No information
	Dichotomized ABCD2 score with DWI		25/479	Stroke	0.81 (0.74-0.88)	No information
Kamouchi 2012 [170]	Fukuoka stroke risk score	Overall ischaemic stroke population	175/3,067	Recurrent ischaemic stroke	0.636 (0.573-0.698)	Hosmer-Lemeshow: x^2=2.30, p=0.97
	Fukuoka stroke risk score	Non-cardiometabolic sub-population	175/3,067	Recurrent ischaemic stroke	0.639 (0.589 - 0.689)	Hosmer-Lemeshow: x^2=8.22, p=0.41
Kernan 2000 [171]	The Stroke Prognosis Instrument II		1331/9678	Stroke or death *(all-cause mortality)*	0.63 (0.62 - 0.65)	No information
Ling 2018 [172]	Modified Essen Stroke Risk Score		85/773	Recurrent ischaemic stroke	0.70 (0.63 - 0.76)	No information

Reference	Model	Cohort name	Sample size (cases/total)	Outcome	C-statistic (95% CI)	Calibration
Rothwell 2005 [178]	ABCD score		Development: 18/209 Validation: 20/190	Stroke	0.85 (0.78-0.91)	No information
Sumi 2013 [185]	Modified Essen stroke Risk Score		121/3452	Recurrent ischemic stroke	0.632 (0.579-0.684)	Hosmer Lemeshow: 8.46 (p = 0.076)
Sumi 2013 [185]	Modified Essen stroke Risk Score		133/3452	CV event (composite of fatal/nonfatal stroke, MI, non-fatal unstable angina, and cardiac death)	0.640 (0.590-0.689)	Hosmer Lemeshow: 7.65 (p = 0.106)
Wijk 2005 [191]	Model 1		1489/2473	All-cause mortality	0.81 (0.79-0.83)	No information
	Model 2		1489/2473		0.82 (0.80-0.83)	No information
	Model 3		1489/2473		0.83 (0.81-0.84)	No information
	Model 1		1336/2473	Composite of death from all vascular causes, non-fatal stroke, and non-fatal MI	0.70 (0.68-0.72)	No information
	Model 2		1336/2473		0.70 (0.68-0.72)	No information
	Model 3		1336/2473		0.72 (0.70-0.74)	No information

CI, confidence interval

3.4.1.3 Risk of bias assessment

PROBAST was used to assess the risk of bias of all eligible studies. Figure 3.2 shows a summary of the risk of bias assessment of models by disease domain. All the model developments had a high risk of bias.

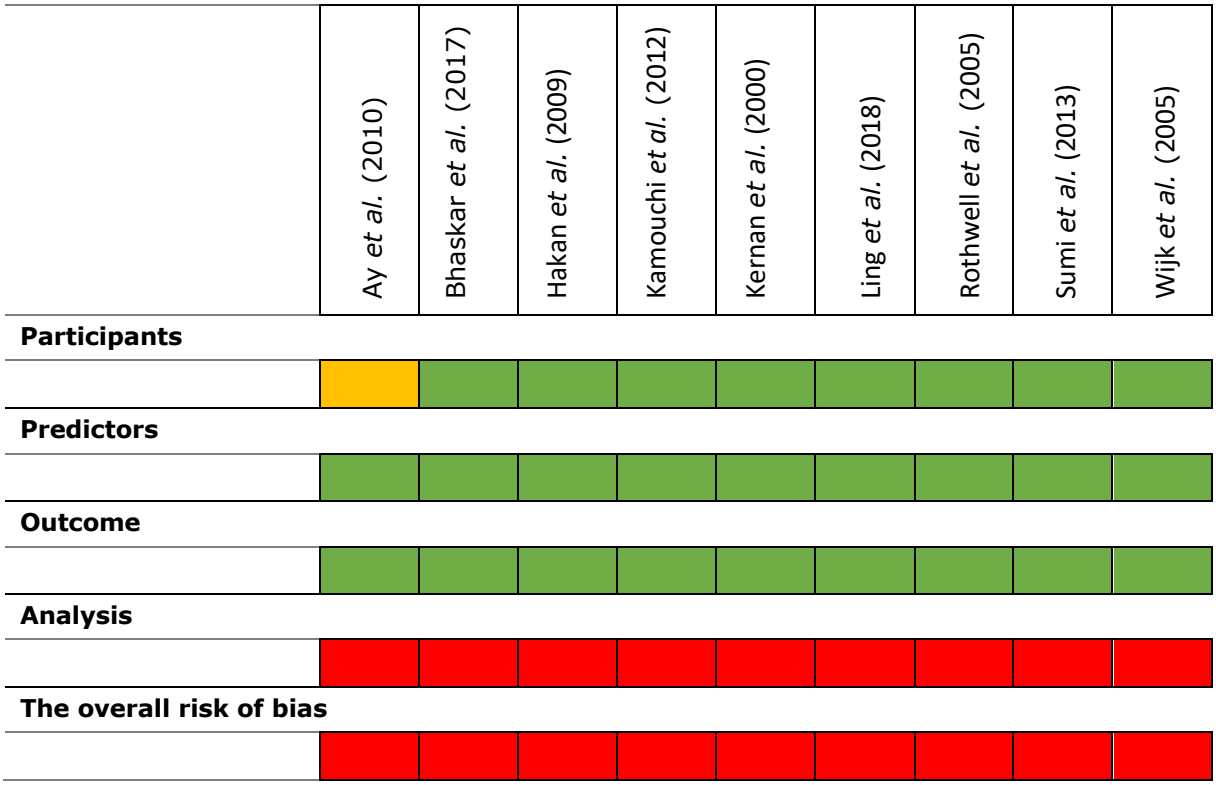


Figure 3.2 Risk of bias assessment

The risk of bias for the studies developing prognostic models was assessed across the four domains: participants, predictors, outcome, and analysis. Each of the domains was judged with "low risk" (depicted in green), "high risk" (red), and "unknown risk of bias" (yellow).

3.4.1.4 Risk factors used in the prognostic models

The number of risk factors (predictors) for the prognostic models ranged between 2 (Fukuoka stroke risk score[170]) and 11 (Modified Essen stroke risk score[172,185]). The most consistent risk factors for the prognostic models assessed were age, sex, blood pressure/hypertension, history of TIA or stroke, and diabetes. Table 3.4 details the risk factors used in the various prognostic models.

Table 3.4 Risk prediction models for stroke and their risk factors

Model	Eligible studies	Risk factors
ABCD score	Bray 2007, Fothergill 2009, Johnston 2007, Purroy 2012, Rothwell 2005, Sciolla 2008	Age, blood pressure, clinical features (weakness/speech disturbance/other symptoms), and duration of symptoms
ABCDI score	Purroy 2012, Sciolla 2008	Age, blood pressure, clinical features, and duration of symptoms and imaging (CT scan findings)
$ABCD_2$ score	Asimos 2010, Chandratheva 2010, Chandratheva 2011, Chatzikonstantinou 2013, Coutts 2008, Fothergill 2009, Ghia 2012, Hakan 2009, Johnston 2007, Maier 2013, Purroy 2012, Sanders 2011, Sheehan 2009, Sheehan 2010, song 2013, Song 2015, Tsivgoulis 2010, Yang 2010	Age, blood pressure, clinical features, duration of symptoms, and diabetes
$ABCD_2I$ score	Purroy 2012	Age, blood pressure, clinical features, duration of symptoms, diabetes, and imaging finding on brain infarction
$ABCD_2$ + MRI	Coutts 2008	Age, blood pressure, clinical features, duration of symptoms, diabetes and MRI of the brain and vasculature findings
$ABCD_3$ score	Purroy 2012	Age, blood pressure, clinical features, duration of symptoms, and prior TIA within 1 week of the index event (dual TIA)
$ABCD_3I$ score	Chatzikonstantinou 2013, Song 2013s	Age, blood pressure, clinical features, duration of symptoms, dual TIA, and ≥50% stenosis on carotid imaging (abnormal DWI)
$ABCD_3V$ score	Purroy 2012	Age, blood pressure, clinical features, duration of symptoms, prior TIA within 1 week of the index event, and vascular imaging information
California risk score	Johnston 2007	Age, clinical features (unilateral weakness, speech impairment), duration of symptoms, and diabetes
CHA_2DS_2-VASc score	Andersen 2015	Age, sex, congestive heart failure, hypertension, stroke/TIA/thromboembolism history, vascular disease history (prior MI, PVD, or aortic plaque), diabetes
Clinical- and Imaging-based predictive (CIP) model	Hakan 2009	TIA with positive diffuse-weighted MRI finding of cerebral infarction in a clinically relevant location

Model	Eligible studies	Risk factors
Essen stroke risk score	Andersen 2015, Ay 2010, Chandratheva 2011, Chen 2016, Ling 2018, Liu 2013, Lui 2017, Maier 2013, Meng 2011, Weimar 2008, Weimar 2009, Weimar 2010, Weimar 2012	Age, hypertension, diabetes, previous myocardial infarction, other cardiovascular diseases (except myocardial infarction and AF), PVD, smoking, and previous stroke or TIA
Fukuoka stroke risk score	Kamouchi 2012	Hypertension and diabetes
Hankey score	Weimar 2010	Age, sex, affected region frequency of TIA, PVD, left ventricular hypertrophy, and residual neurological signs
Life-Long After Cerebral Ischemia Trial (LiLAC) score	Weimar 2010	TIA, minor ischemic stroke
Model 1	Bhaskar 2017	Cerebral infarction, stroke not specified as haemorrhage or infarction
Model 1	Wijk 2005	Age, sex, medical history, and current drug use
Model 2	Wijk 2005	Age, sex, medical history, current drug use, neurological symptoms, and examination findings
Model 3	Wijk 2005	ECG and CT-scan findings
Modified Essien stroke risk score	Ling 2018, Sumi 2013	Age, hypertension, diabetes, previous myocardial infarction, other cardiovascular diseases, PVD, smoking, previous TIA/stroke, stroke subtype, waist circumference, sex
Recurrent Risk Estimator score	Arsava 2011, Arsava 2016, Ay 2010, Maier 2013, Song 2015	History of TIA/stroke within the month preceding index stroke, admission stroke subtype according to Causative Classification of Stroke System (CCS), and MRI imaging findings—isolated cortical infarcts, multiple acute infarcts, simultaneous infarcts in different vascular territories, and multiple infarcts of different ages
Stroke Prognosis Instrument I	Kernan 2000	Age, diabetes, and hypertension
Stroke Prognosis Instrument II	Ay 2010, Chandratheva 2011, Kernan 2000, Liu 2013, Meng 2011, Weimar 2010, Weimar 2012	Age, diabetes, hypertension, congestive heart failure, and prior stroke

3.5 Discussion

This systematic review identified 23 prognostic models predicting subsequent major adverse cardiovascular events (MACE) in an adult with an established diagnosis of stroke. There were 11 prognostic model developments and 75 external validations of models in 40 eligible studies identified. All the models developed were assessed as being at high risk of bias due to limitations with the modelling methodology. Relatively, the ABCD and Recurrent Risk Estimator prognostic models had the best predictive accuracy (discrimination).

Most models developed and validated externally in this systematic review were from developed countries such as the USA, UK, China, and the Netherlands. Age, sex, history of TIA, hypertension (blood pressure), and diabetes were more frequently used as predictors in the models for predicting subsequent cardiovascular morbidity or mortality outcomes in adult patients with stroke. None of these frequently used predictors was unexpected. For instance, older age is an independent risk factor for recurrent stroke, CHD, and death in patients with stroke.[193] Also, prior history of hypertension is a known risk factor for a recurrent cardiovascular event.[194]

The review highlights methodological limitations in the development of the prognostic models resulting in a high risk of bias. In particular, improper or no information on the handling of missing data, selection of candidate predictor variables, or incomplete evaluation of model performance (no model calibration). Reviews of prognostic models developed for other diseases, including oropharyngeal cancer, chronic lymphocytic leukaemia, and chronic obstructive pulmonary disease,[195–198] also found inadequate sample size, improper handling of missing data, and incomplete evaluation of model performance. Research should pay attention to each detail in the modelling process to get prognostic models with good predictive capabilities for clinical practice.

The ABCD and Recurrent Risk Estimator prognostic models were the best performing models, hence could be considered to be the most useful models for predicting those stroke patients most likely to have major adverse clinical outcomes. However, these prognostic models need to be externally validated in diverse large population cohorts to ensure they are clinically fit for use in clinical practice.

Strengths and limitations

This review has a number of strengths. This is the first systematic review of prognostic prediction models for MACE outcome in adults with an established diagnosis of stroke. The most recent guideline of systematic reviews for risk prediction models[199] and the PRISMA-P 2015 statement[83] was used to guide this review. A formal and broad search strategy was employed without language restrictions. With a robust tool such as the Prediction Model Study Risk Of Bias Assessment Tool (PROBAST), the risk of bias for included studies was assessed.[151]

This systematic review also has certain limitations. The data sources used in the respective studies were derived from a variety of clinical settings, both within and across different countries with different patient characteristics and clinical practices. Despite the clinical heterogeneity within included studies, conclusions can be drawn from across the diverse settings. There were, however, different time points and settings for reporting composite MACE outcome or any of its components, making meta-analysis a challenge.

Summary

This systematic review identified 40 peer-reviewed articles reporting 23 distinct prognostic models (11 model developments and 75 external validations) for predicting subsequent major adverse cardiovascular events (MACE) in an adult with an established diagnosis of stroke. The risk of bias was high for all the 11

prognostic model developments. Since accurate risk scores may provide additional options for stratifying patients with stroke early, thus lowering the burden of this disease, their further development, validation, and implementation should be a research and public health priority.

Chapter 4

Sex, age, and socioeconomic differences in non-fatal stroke incidence and subsequent major adverse outcomes

The previous chapter presented the findings of a systematic review that explored prognostic models predicting MACE outcomes in adult patients with a diagnosis of stroke. This chapter explores the age, sex, and socioeconomic differences in the rates of incident non-fatal stroke and first subsequent MACE outcomes using linked datasets.

A paper based on this research study has been published in the journal *Stroke*:

Akyea, R. K., Vinogradova, Y., Qureshi, N., Patel, R. S., Kontopantelis, E., Ntaios, G., Asselbergs F.W., Kai J., Weng, S. F. (2021). Sex, Age, and Socioeconomic Differences in Nonfatal Stroke Incidence and Subsequent Major Adverse Outcomes. Stroke, 52(2), 396-405. https://doi.org/10.1161/strokeaha.120.031659

4.1 Abstract

Background: Data about variations in stroke incidence and subsequent major adverse outcomes are essential to inform secondary prevention and prioritising resources to those at greatest risk of major adverse endpoints. This chapter describes the age, sex, and socioeconomic differences in the rates of first non-fatal stroke and subsequent major adverse outcomes.

Methods: The cohort study used linked CPRD GOLD and HES APC data. The incidence rate ratio (IRR) of first non-fatal stroke and subsequent major adverse outcomes (MACE, recurrent stroke, CVD-related and all-cause mortality) were calculated and presented by year, sex, age group, and socioeconomic status (SES) based on an individual's location of residence, in adults with incident non-fatal stroke diagnosis between 1998 and 2017.

Results: There were 82,774 incident non-fatal stroke events recorded in either primary care or hospital data – an incidence rate of 109.20 per 100,000 person-years (95% CI: 108.46–109.95). The incidence of non-fatal stroke was significantly higher in women when compared with men (IRR 1.13, 95% CI: 1.12–1.15; $p<0.001$). Rates adjusted for age and sex were higher in the lowest compared to the highest SES group (IRR 1.10, 95% CI: 1.08–1.13, $p<0.001$).

For subsequent major adverse outcomes, the overall incidence for MACE was 38.05 per 100 person-years (95% CI: 37.71–38.39) with a slightly higher incidence in women compared to men (38.42 vs 37.62; IRR 1.02, 95% CI: 1.00–1.04, $p=0.0229$). Age and SES largely accounted for the observed higher incidence of adverse outcomes in women.

Conclusions: In the UK, the incidence of initial non-fatal stroke and subsequent major adverse outcomes are higher in women, older populations, and people living in socially deprived areas.

4.2 Introduction

The ageing population and treatment improvements are expected to significantly impact stroke epidemiology[200] – the number of stroke survivors is on the rise, impacting the need for post-stroke facilities and the risk of stroke recurrence.[201] About 1 in 4 stroke survivors will experience another stroke within 5 years.[144] Despite advances in the management of stroke patients, mortality and disability rates remain high.[202]

There is a lack of contemporary evidence from nationally representative data across the entire spectrum of both primary and secondary care, to assess demographic variations in the incidence of first non-fatal stroke. Most published estimates of stroke incidence in the UK only capture certain types of stroke, may or may not include recurrent stroke, focus on small non-representative populations or a short period.[203–205] Moreover, the sex- and age- variations for the incidence of major adverse outcomes after the first stroke have either been in selected populations with incident ischaemic stroke and/or above a pre-defined age.[206,207] To identify patterns and any discrepancies in the care of patients, information on temporal trends in sex- and age-related differences with respect to the incidence of stroke and subsequent major adverse outcomes are needed. This can inform resource allocation and policy to enhance clinical care and outcomes.

In this study, I sought to update current knowledge on differences in non-fatal stroke incidence and incidence of major adverse outcomes following the first non-fatal stroke event. I used linked electronic health records from primary care consultations, secondary care (hospital admissions and procedure-level data), and the national death registry that are representative of the UK population. This population-based study explored demographic variations in the incidence of stroke (first ever non-fatal stroke) and incidence of major adverse outcomes after first-ever stroke among individuals aged 18 years and over.

4.3 Methods

Data source

This prospective cohort study used the UK Clinical Practice Research Datalink (CPRD GOLD) of primary care electronic health records,[87] linked to Hospital Episode Statistics (HES APC),[208] Office for National Statistics (ONS) mortality data,[95] and social deprivation data.[97] The databases have been previously described in Chapter 2 (Section 2.2).

Study population

The study cohort of patients with the first record of non-fatal stroke in either CPRD GOLD or HES APC between 1 January 1998 and 31 December 2017 has been previously described in Chapter 2 (Section 2.3.1).

Identifying patients, patient characteristics, and outcomes

To identify patients with stroke and outcomes of interest, the stroke code lists from the CALIBER code repository[100] were used – from CPRD GOLD using Read codes, from HES APC using ICD-10 codes detailed in Chapter 2 (Section 2.3.2.1). I also extracted information on demographic factors including age, sex, SES, and ethnicity. Socioeconomic status based on the English Index of Multiple Deprivation (IMD) 2015,[97] described in Chapter 2 (Section 2.2.4), was categorised into quintiles (quintile 1 – highest SES group to quintile 5 – lowest SES group).

The outcomes of interest were subsequent major adverse events (composite major adverse cardiovascular events (MACE), recurrent stroke, cardiovascular (CVD)-related and all-cause mortality) after incident stroke. MACE was defined as a composite of CHD, stroke, PVD, heart failure, and CVD-related death) based on records from CPRD GOLD, HES APC, or ONS registries.

Statistical analysis

Baseline characteristics were presented by sex and expressed using mean and standard deviation for continuous variables (after assessment of the normality of data) and percentages for categorical variables. Differences for categorical and continuous variables were assessed using Chi-squared and t-tests respectively. Incidence rates per 100,000 person-years for stroke were calculated by dividing the number of individuals with a stroke by the number of person-years of all patients in the original cohort. Incidence rates were presented by age (5-year intervals), sex, and year of diagnosis. Incidence rate ratios for IMD quintiles were adjusted for age and sex in a Poisson regression model.

I calculated the incidence rates per 100 person-years for subsequent major adverse outcomes in the cohort of patients with no prior history of a major adverse event. A sensitivity analysis was restricted to the cohort of patients with subsequent major adverse outcomes occurring after 30 days of the index stroke was done. Patients with other associated major adverse endpoints are more likely to remain in the hospital for an extended period following admission for their incident stroke event. The recording of such hospital (secondary care) activity in primary care records through discharge letters or referral notes from specialists may be delayed. The date of referral/letter may erroneously be recorded as the date of stroke or associate major adverse outcome. Some outcome events may only be recorded when a post-hospitalisation visit to primary care occurs. Hence, a 30-day interval was chosen because records for adverse outcomes occurring within 30 days is likely to be related to index stroke event.[209,210]

The study findings are reported following the Reporting of Studies Conducted Using Observational Routinely Collected Health Data (RECORD) recommendations.[211] All statistical analyses were performed using Stata 16.1 (StataCorp LP) and statistical significance was set at $p<0.05$. Figure 4.1 illustrated the study flow.

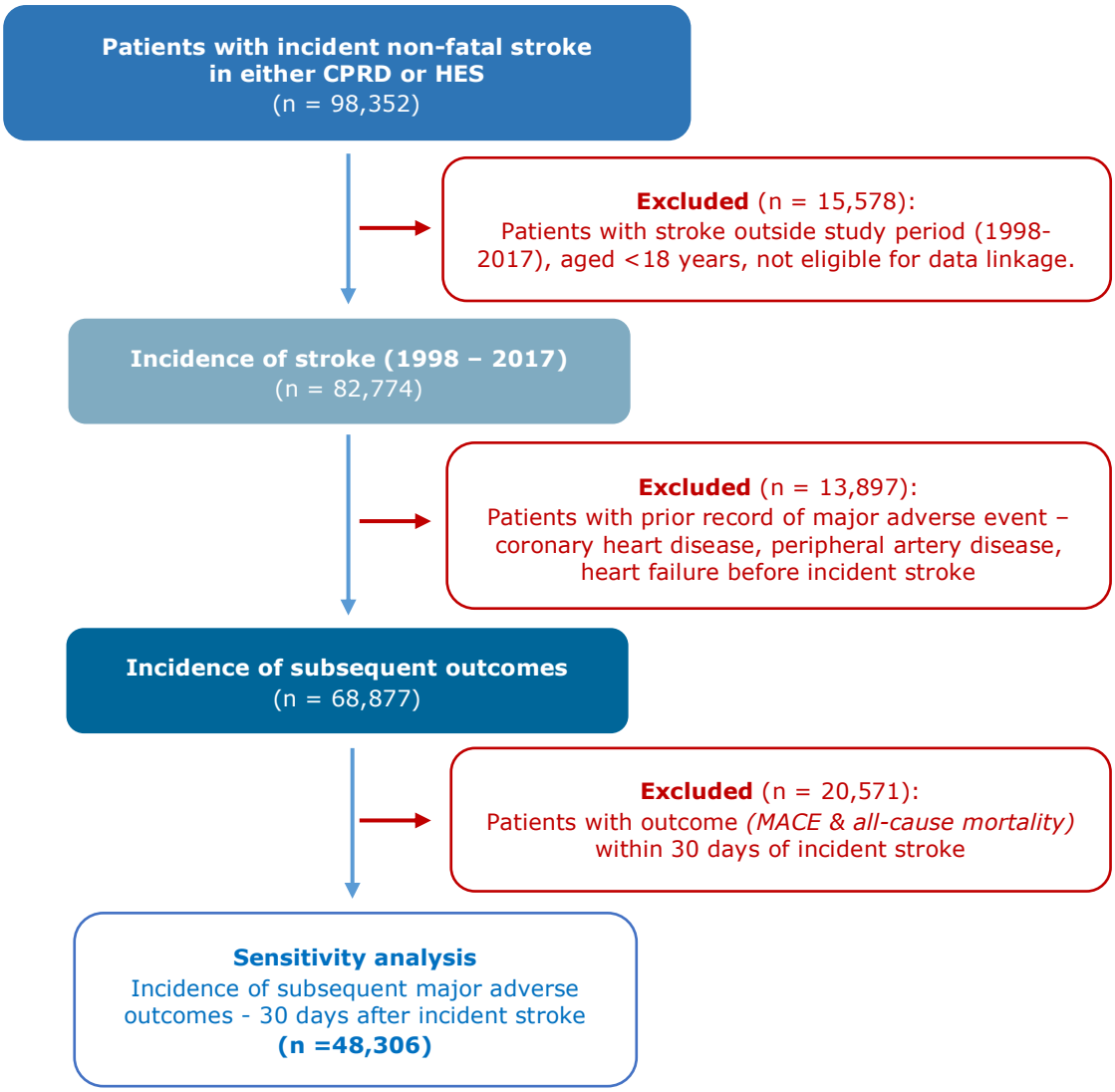


Figure 4.1 Study flow diagram

4.4 Results

A total of 9,992,380 individuals 18 years and over were identified in CPRD GOLD with a total follow-up time of 75,794,468.8 person-years between 1 January 1998 and 31 December 2017. The mean follow-up time was 1.81 years (standard deviation (SD): 2.78) with a median of 0.51 years (interquartile range (IQR): 0.05 – 2.41). There were 82,774 individuals with incident non-fatal stroke. There were 44,614 (53.9%) women. The mean age for incident stroke was 74.3 years (SD: 13.6). Males had an incident stroke at a younger age compared to women (71.4

vs 76.9 years). Hypertension was the most prevalent comorbid condition (48.4%) – see Table 4.1 for details.

Table 4.1 Demographic characteristics of individuals aged 18 years or above with incident non-fatal stroke (n=82,774)

	Total	**Men**	**Women**	***p*-value**
	n = 82,774	n = 38,160 (46.1%)	n = 44,614 (53.9%)	
Age at incident stroke	74.3 (13.6)	71.4 (13.2)	76.9 (13.5)	0.0001
Type of incident stroke				<0.001
Haemorrhagic	7,855 (9.5)	3,842 (10.1)	4,013 (9.0)	
Ischaemic	31,777 (38.4)	15,151 (39.7)	16,626 (37.3)	
Not otherwise specified	43,142 (52.1)	19,167 (50.2)	23,975 (53.7)	
Socioeconomic status				0.170
1 (Highest SES)	17,491 (21.1)	8,209 (21.5)	9,282 (20.8)	
2	18,288 (22.1)	8,359 (21.9)	9,929 (22.3)	
3	17,923 (21.7)	8,259 (21.6)	9,664 (21.7)	
4	15,356 (18.6)	7,001 (18.4)	8,355 (18.7)	
5 (Lowest SES)	13,569 (16.4)	6,262 (16.4)	7,307 (16.4)	
Missing	147 (0.2)	70 (0.2)	77 (0.2)	
Ethnicity				<0.001
Asian	1,063 (1.3)	579 (1.5)	484 (1.1)	
Black	633 (0.8)	317 (0.8)	316 (0.7)	
Mixed	116 (0.1)	64 (0.2)	52 (0.1)	
Other	581 (0.7)	294 (0.8)	287 (0.6)	
White	73,764 (89.1)	34,157 (89.5)	39,607 (88.8)	
Unknown	6,617 (8.0)	2,749 (7.2)	3,868 (8.7)	
Comorbid conditions				
Atrial fibrillation	10,316 (12.5)	4,446 (11.7)	5,870 (13.2)	<0.001
Diabetes mellitus	11,014 (13.3)	5,630 (14.8)	5,384 (12.1)	<0.001
Dyslipidaemia	8,892 (10.7)	3,894 (10.2)	4,998 (11.2)	<0.001
Hypertension	40,411 (48.8)	17,238 (45.2)	23,173 (51.9)	<0.001
TIA	17,365 (21.0)	7,790 (20.4)	9,575 (21.5)	<0.001

TIA: transient ischaemic attack; n: total number; %: percentage/proportion; Mean age at incident stroke reported with standard deviation.

Overall stroke incidence

The overall incidence rate (IR) of stroke from 1998 to 2017 was 109.21 per 100,000 person-years (95% CI: 108.47–109.96). The incidence rate was relatively steady between 1998 and 2003, with a peak incidence in 2004, and a subsequent decline in incidence till 2017, as shown in Figure 4.2 (details in Appendix D.4.1). The overall stroke incidence was higher in women (IR: 115.84, 95% CI: 114.77–116.92) compared to males (IR: 102.36, 95% CI: 101.34–103.39) with an incidence rate ratio (IRR) of 1.13 (95% CI: 1.12–1.15, p<0.0001). Stroke incidence increased in age groups older than the 55-59-year group – Figure 4.3 show variations and Appendix D.4.2 provides the detailed results. Males aged 30-74 years had higher stroke incidence rates compared to women, however, from age 75 the incidence rates were much higher in women.

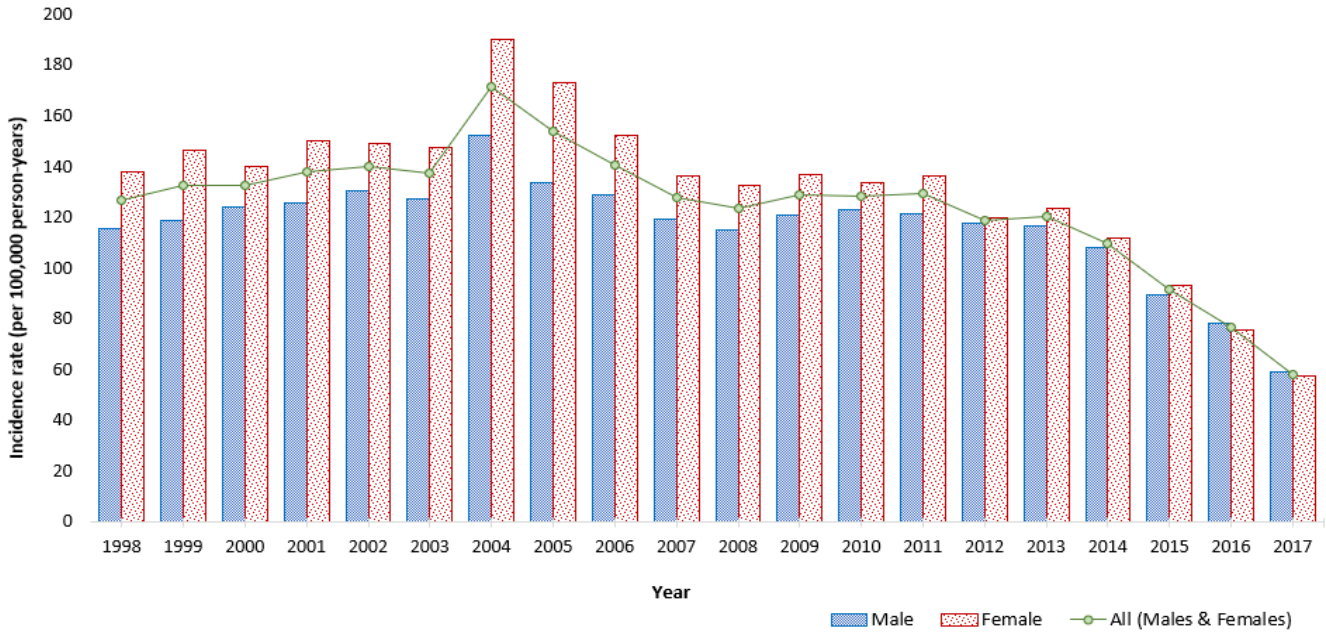


Figure 4.2 Trends in stroke incidence by sex (1998 – 2017)

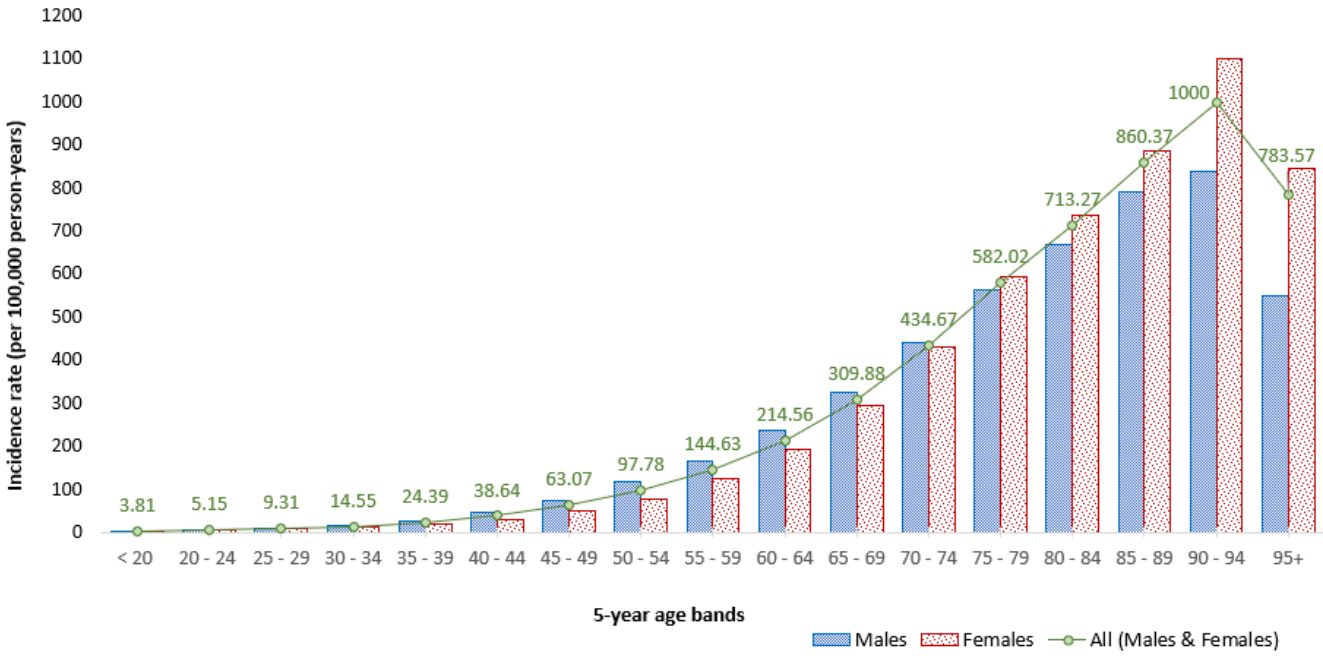


Figure 4.3 Trends in stroke incidence by age group and sex (1998 – 2017)

Stroke incidence by socioeconomic status

After adjusting for the effects of age and sex (Table 4.2), for every increase in IMD quintile, the incidence of stroke increased. The rate of stroke incidence among individuals in the lowest SES quintile was 10% higher than the rate in the highest SES quintile (IRR, 1.10, 95% CI 1.08 – 1.13).

Table 4.2 Age- and sex-adjusted incidence rate ratio of stroke, by socioeconomic status

Socioeconomic status	Incident stroke
1 (least deprived)	Reference
2	1.03 (1.01–1.05)
3	1.04 (1.01–1.06)
4	1.09 (1.06–1.11)
5 (most deprived)	1.10 (1.08–1.13)

Subsequent major adverse outcomes

Of the 82,774 individuals with incident stroke events, 13,897 had a prior history of major adverse outcomes and were excluded. Of the 68,877 individuals, the mean age for incident stroke was 73.3 years (SD: 13.9), with 37,395 (54.3%) being women. With respect to outcomes, 47,500 (69.0%) had a MACE; 33,831 (49.1%) recurrent strokes [haemorrhagic stroke: 2,378 (4.1%), ischaemic stroke: 8,842 (15.1%), stroke (not specified): 22,611 (38.6)]; 9,174 (13.3%) cardiovascular death; and 20,335 (29.5%) all-cause mortalities, occurring after the incident stroke events – Table 4.3. There were 25,731 (68.8%) women with MACE outcomes. Figure 4.4 shows the distribution of MACE, recurrent stroke, cardiovascular and all-cause mortality outcomes presented by sex and across 5-year age bands. Most subsequent outcomes occurred within 2 years of incident stroke – with the median follow-up time at which outcomes occurred after incident stroke ranging between 0.10 years (IQR: 0.02 – 1.49) for CVD-related mortality and 1.74 years (IQR: 0.51 – 4.42) for heart failure.

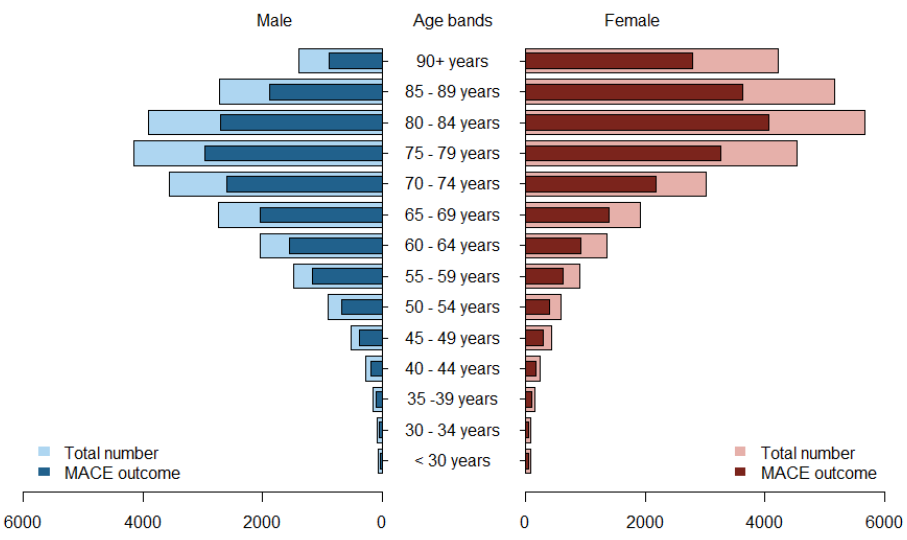
Table 4.3 Demographic characteristics of individuals aged 18 years or above with incident non-fatal stroke and no prior history of major adverse event (n = 68,877)

	Total n = 68,877	**Men** n = 31,482 (45.7%)	**Women** n = 37,395 (54.3%)	***p*-value**
Age at incident stroke	73.3 (13.9)	70.3 (13.4)	75.9 (13.9)	0.0001
Type of incident stroke				<0.001
Haemorrhagic	6,682 (9.7)	3,229 (10.3)	3,453 (9.2)	
Ischaemic	26,146 (38.0)	12,391 (39.4)	13,755 (36.8)	
Not otherwise specified	36,049 (52.3)	15,862 (50.4)	20,187 (54.0)	
Socioeconomic status				0.282
1 (Highest SES)	14,779 (21.5)	6,840 (21.7)	7,939 (21.2)	
2	15,350 (22.3)	6,934 (22.0)	8,416 (22.5)	
3	14,870 (21.6)	6,782 (21.5)	8,088 (21.6)	
4	12,661 (18.4)	5,748 (18.3)	6,913 (18.5)	
5 (Lowest SES)	11,101 (16.1)	5,119 (16.3)	5,982 (16.0)	
Missing	116 (0.2)	59 (0.2)	57 (0.2)	
Ethnicity				<0.001
Asian	895 (1.3)	479 (1.5)	416 (1.1)	
Black	560 (0.8)	278 (0.9)	282 (0.8)	
Mixed	102 (0.2)	57 (0.2)	45 (0.1)	
Other	481 (0.7)	243 (0.8)	238 (0.6)	
White	61,145 (88.8)	28,070 (89.2)	33,075 (88.5)	
Unknown	5,694 (8.3)	2,355 (7.5)	3,339 (8.9)	
Comorbid conditions				
Atrial fibrillation	6,456 (9.4)	2,746 (8.7)	3,710 (9.9)	<0.001
Diabetes mellitus	7,979 (11.6)	3,968 (12.6)	4,011 (10.7)	<0.001
Dyslipidaemia	6,562 (9.5)	2,809 (8.9)	3,753 (10.0)	<0.001
Hypertension	31,861 (46.3)	13,388 (42.5)	18,473 (49.4)	<0.001
TIA	14,073 (20.4)	6,257 (19.9)	7,816 (20.9)	0.001
Major adverse outcomes				<0.001
Coronary heart disease	2,420 (4.1)	1,311 (5.0)	1,109 (3.5)	
Haemorrhagic stroke	2,378 (4.1)	1,209 (4.6)	1,169 (3.6)	
Ischaemic stroke	8,842 (15.1)	4,254 (16.1)	4,588 (14.3)	
Stroke (not specified)	22,611 (38.6)	10,512 (39.7)	12,099 (37.6)	
Peripheral vascular disease	593 (1.0)	334 (1.3)	259 (0.8)	
Heart failure	1,482 (2.5)	633 (2.4)	849 (2.6)	
CVD-related death	9,174 (15.6)	3,516 (13.3)	5,658 (17.6)	
Non-CVD related death	11,161 (19.0)	4,734 (17.9)	6,427 (20.0)	

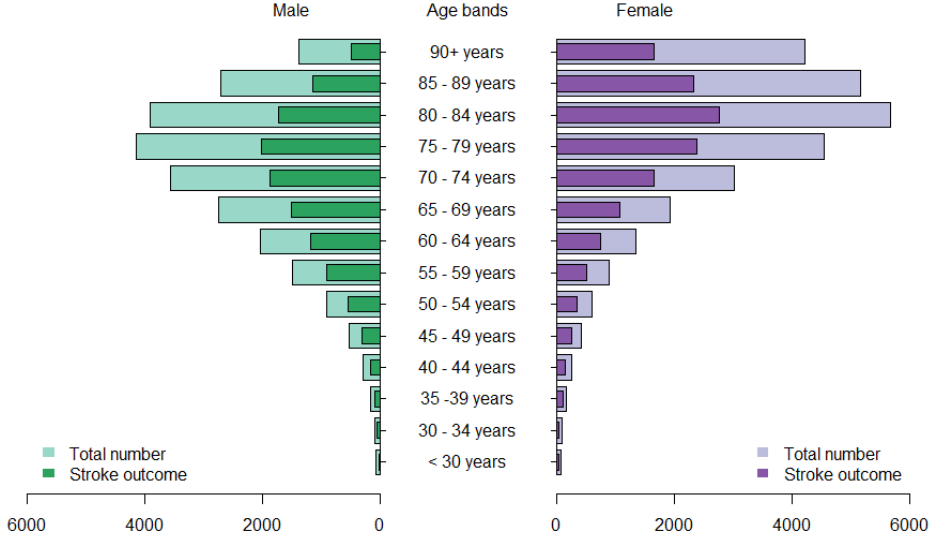
TIA: transient ischaemic attack; n: total number; %: percentage/proportion; Mean age at incident stroke reported with standard deviation.

Major adverse event is defined as a record of either coronary heart disease, peripheral vascular disease, or heart failure.

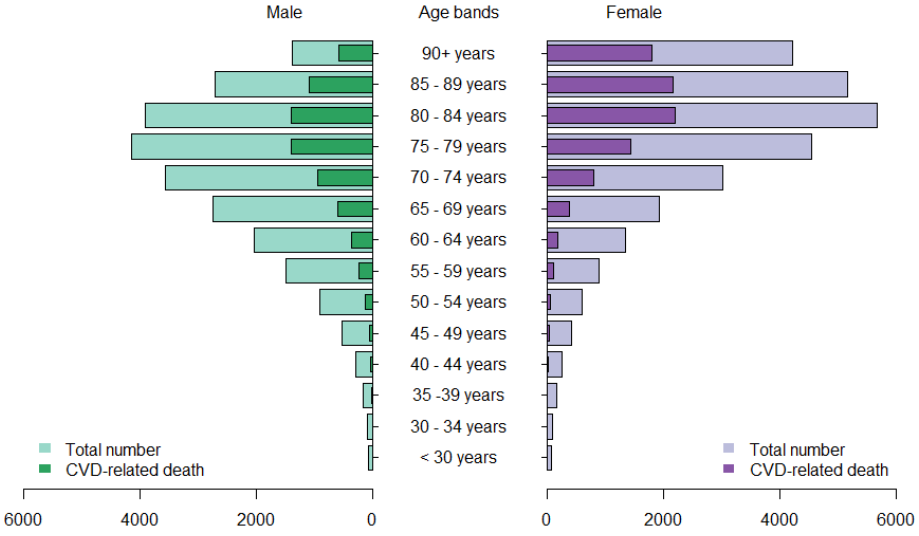
Major adverse cardiovascular events, n=37,082 (70.8%)



Recurrent stroke, n=26,065 (49.8%)



Cardiovascular-related mortality, n=15,974 (30.5%)



All-cause mortality, n=39,320 (75.1%)

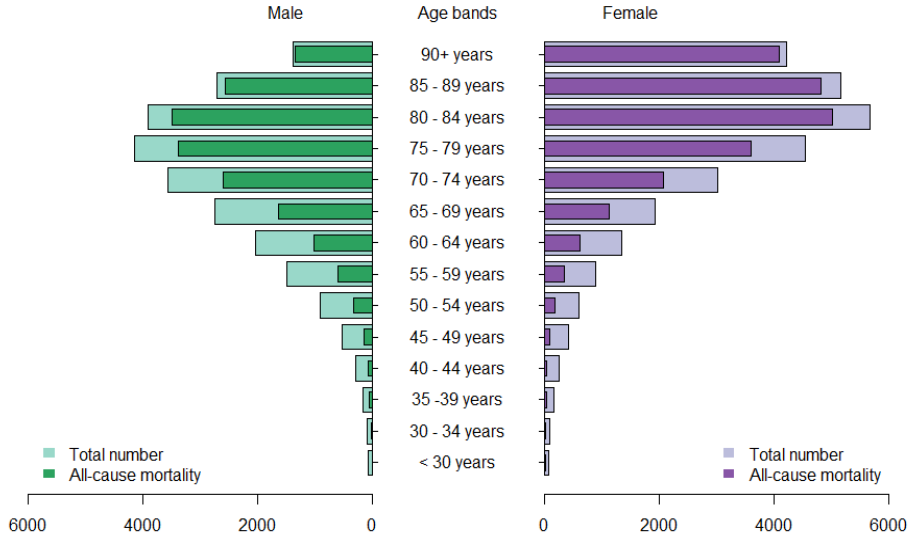
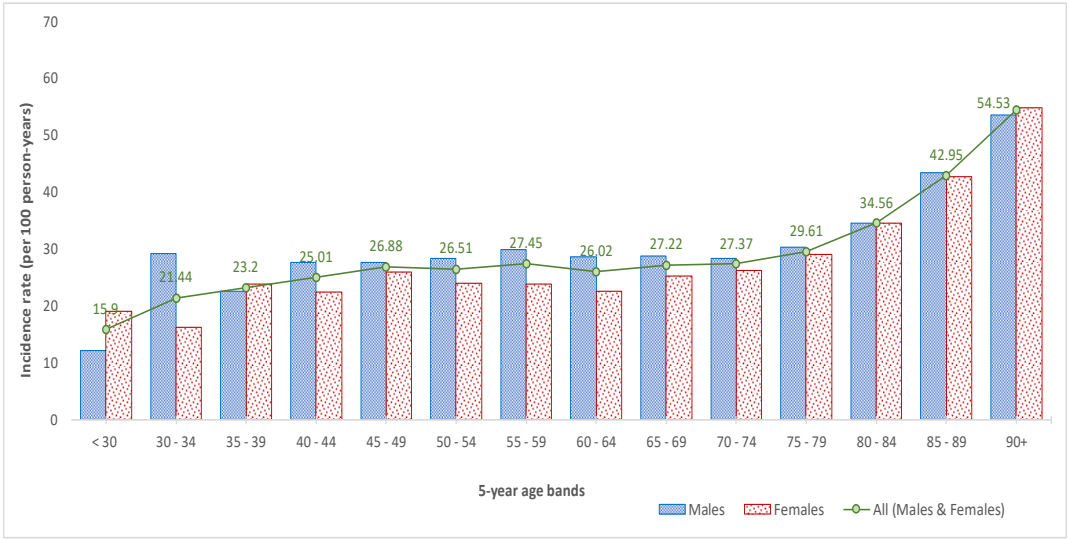


Figure 4.4 Distribution of subsequent major adverse outcomes presented by sex and 5-year age groups (n = 52,362)

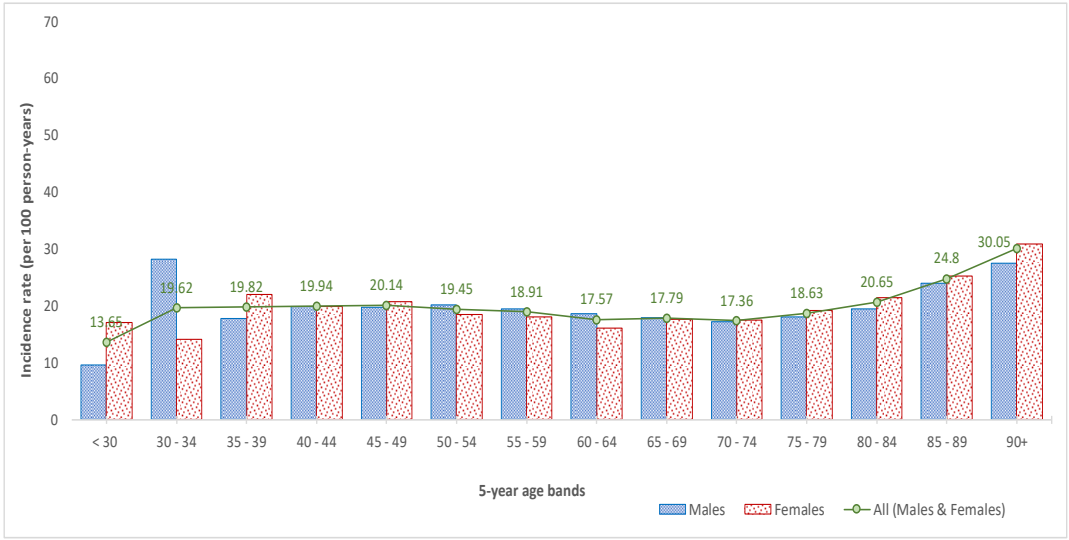
Incidence of subsequent major adverse outcomes

The overall MACE incidence rate was 38.05 per 100 person-years (95% CI: 37.71–38.39). There was a steady rise in MACE incidence across the various age groups before peaking in the 80+ age group with a MACE incidence rate of 45.31 per 100 person-years as illustrated in Figure 4.5. For the constituent MACE outcomes, women had a higher incidence rate for CVD-related (4.13 vs 2.72 per 100 person-years respectively; IRR 1.52, 1.45–1.58) and all-cause mortality (8.45 vs 6.21 per 100 person-years respectively; IRR 1.36, 1.32–1.40). The incidence of coronary heart disease and peripheral vascular disease were, however, higher in men. Table 4.4 details the sex variation in the incidence of the constituent MACE outcomes. In comparing women to men, the age- and SES-adjusted sex-specific IRR for MACE was 0.92 (0.90–0.94), recurrent stroke: 0.96 (0.94–0.98), CVD-related death: 1.00 (0.96–1.05), all-cause mortality: 0.96 (0.93–0.98), CHD: 0.75 (0.69–0.81), PVD: 0.64 (0.54–0.76) and heart failure: 0.93 (0.84–1.03).

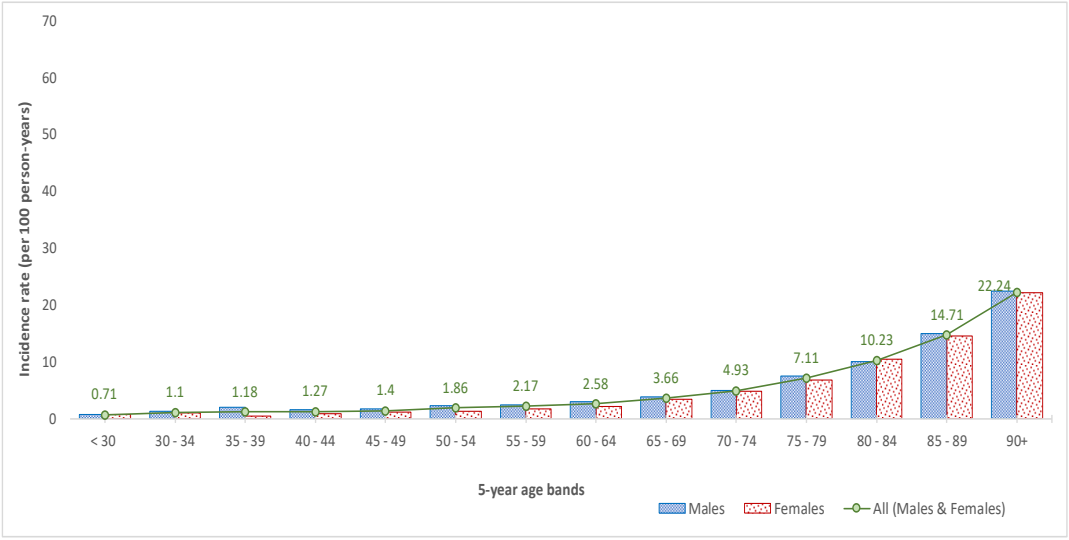
Major adverse cardiovascular events



Recurrent stroke



Cardiovascular-related mortality



All-cause mortality

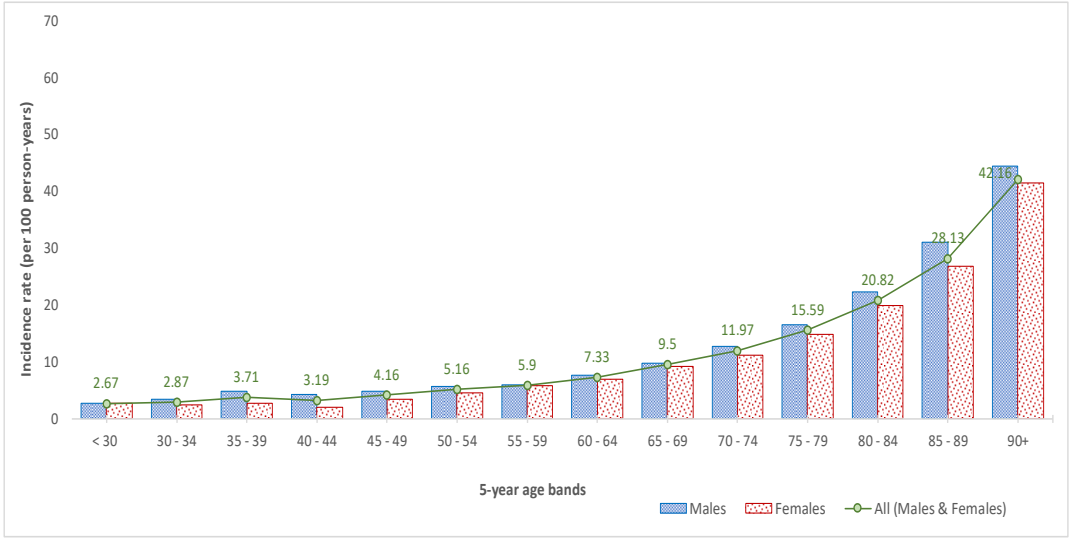


Figure 4.5 Incidence of subsequent major adverse outcomes presented by sex and 5-year age groups (n = 52,362)

Table 4.4 **Incidence of subsequent major adverse outcomes** (n = 68,877)

	Follow-up time (years)	Cases	Person-years*	Incidence rate (per 100 person-years)	Incidence rate ratio	*p*-value
MACE (All)	0.20 (0.03 – 1.51)	47,500	1,200	38.05 (37.71 – 38.39)		0.0229
Men	0.19 (0.03 – 1.49)	21,769	578.65	37.62 (37.12 – 38.12)	Reference	
Women	0.21 (0.03 – 1.51)	25,731	669.78	38.42 (37.95 – 38.89)	1.02 (1.00 – 1.04)	
Recurrent stroke (All)	0.16 (0.03 – 1.20)	33,831	1,300	25.80 (25.53 – 26.08)		0.2582
Men	0.14 (0.03 – 1.11)	15,975	623.23	25.63 (25.24 – 26.03)	Reference	
Women	0.19 (0.04 – 1.29)	17,856	688.09	25.95 (25.57 – 26.33)	1.01 (0.99 – 1.03)	
Coronary heart disease (All)	1,71 (0.44 – 4.12)	2,420	2,600	0.95 (0.91 – 0.99)		<0.0001
Men	1.72 (0.42 – 4.33)	1,311	1,200	1.06 (1.01 – 1.12)	Reference	
Women	1.71 (0.47 – 3.98)	1,109	1,300	0.84 (0.79 – 0.89)	0.79 (0.73 – 0.86)	
Peripheral arterial disease (All)	1.60 (0.55 – 3.94)	593	2,600	0.23 (0.21 – 0.25)		0.0002
Men	1.59 (0.52 – 3.91)	334	1,300	0.26 (0.24 – 0.29)	Reference	
Women	1.60 (0.55 – 3.94)	259	1,300	0.19 (0.17 – 0.22)	0.73 (0.62 – 0.87)	
Heart failure (All)	1.74 (0.51 – 4.42)	1,482	2,600	0.57 (0.54 – 0.60)		<0.0001
Men	1.73 (0.52 – 4.29)	633	1,300	0.50 (0.46 – 0.54)	Reference	
Women	1.74 (0.51 – 4.61)	849	1,300	0.64 (0.60 – 0.68)	1.27 (1.15 – 1.41)	
Cardiovascular mortality (All)	0.10 (0.02 – 1.49)	9,174	2,700	3.44 (3.38 – 3.52)		<0.0001
Men	0.11 (0.02 – 1.79)	3,516	1,300	2.72 (2.63 – 2.81)	Reference	
Women	0.09 (0.02 – 1.27)	5,658	1,400	4.13 (4.02 – 4.24)	1.52 (1.45 – 1.58)	
All-cause mortality (All)	0.55 (0.05 – 3.20)	20,335	2,800	7.37 (7.27 – 7.47)		<0.0001
Men	0.66 (0.06 – 3.39)	8,250	1,300	6.21 (6.07 – 6.34)	Reference	
Women	0.49 (0.05 – 3.10)	12,085	1,400	8.45 (8.30 – 8.60)	1.36 (1.32 – 1.40)	

* 100 person-years at risk; All – both men and women; Follow-up time: median follow-up time in years reported with interquartile range.

Subsequent major adverse outcomes by socioeconomic status

After adjusting for age and sex, the rate ratio of MACE incidence among individuals in the lowest SES quintile was 9% more than the rate in the highest SES quintile (IRR, 1.09, 95% CI 1.06–1.13). There was no significant difference in recurrent stroke incidence between individuals in the lowest and highest SES quintiles, IRR 1.00 (95% CI 0.97–1.04) – Table 4.5.

Table 4.5 Age- and sex-adjusted incidence rate ratio of subsequent major adverse outcomes, by socioeconomic status

Socioeconomic status	MACE	Recurrent stroke	CVD-related mortality	All-cause mortality
1 (Highest SES)	Reference	Reference	Reference	Reference
2	1.04 (1.01 – 1.07)	1.04 (1.00 – 1.07)	1.10 (1.03 – 1.17)	1.07 (1.02 – 1.11)
3	1.09 (1.06 – 1.12)	1.07 (1.04 – 1.11)	1.07 (1.00 – 1.14)	1.05 (1.00 – 1.09)
4	1.11 (1.08 – 1.14)	1.05 (1.01 – 1.08)	1.21 (1.13 – 1.29)	1.16 (1.11 – 1.21)
5 (Lowest SES)	1.09 (1.06 – 1.13)	1.00 (0.97 – 1.04)	1.31 (1.23 – 1.41)	1.31 (1.26 – 1.37)

MACE: major adverse cardiovascular event

Sensitivity analysis

For the sensitivity analysis, 20,571 (29.9%) of the 68,877 patients with the subsequent major adverse outcome with 30days of incident stroke were excluded. The remaining follow-up cohort (n=48,306) had similar strata for SES as the excluded cohort. The proportion of patients with pre-stroke comorbid conditions varied between the excluded and remaining cohorts: atrial fibrillation (10.9% vs 8.7%, $p<0.001$), dyslipidaemia (8.3% vs 10.0%, $p<0.001$), transient ischaemia attack (8.2% vs 25.6%, $p<0.001$) respectively – Table 4.6.

Table 4.6 Descriptive characteristics of patients with a subsequent outcome within 30 days compared to those with outcomes after 30 days of incident stroke

	Outcome within 30 days n=20,571 (29.9%)	**Outcome after 30 days** n=48,306 (70.1%)	***p*-value**
Age at incident stroke	73.8 (14.0)	73.1 (13.9)	0.0001
Age at subsequent MACE outcome	73.6 (14.0)	76.3 (13.1)	0.0001
Female	10,995 (53.5)	26,400 (54.7)	0.004
Socioeconomic status			0.714
1 (Highest SES)	4,457 (21.7)	10,322 (21.4)	
2	4,561 (22.2)	10,789 (22.3)	
3	4,481 (21.8)	10,389 (21.5)	
4	3,783 (18.4)	8,878 (18.4)	
5 (Lowest SES)	3,254 (15.8)	7,847 (16.2)	
Missing	35 (0.2)	81 (0.2)	
Ethnicity			<0.001
Asian	280 (1.4)	615 (1.3)	
Black	181 (0.9)	379 (0.8)	
Mixed	29 (0.1)	73 (0.2)	
Other	145 (0.7)	336 (0.7)	
White	17,966 (87.3)	43,179 (83.4)	
Unknown	1,970 (9.6)	3,724 (7.7)	
Comorbid conditions			
Atrial fibrillation	2,244 (10.9)	4,212 (8.7)	<0.001
Diabetes mellitus	2,484 (12.1)	5,495 (11.4)	0.009
Dyslipidaemia	1,715 (8.3)	4,847 (10.0)	<0.001
Hypertension	9,400 (45.70)	22,461 (46.5)	0.053
TIA	1,696 (8.2)	12,377 (25.6)	<0.001
Major adverse outcomes			<0.001
Coronary heart disease	291 (1.4)	2,129 (5.6)	
Haemorrhagic stroke	1,448 (7.0)	930 (2.44)	
Ischaemic stroke	5,247 (25.5)	3,595 (9.4)	
Stroke (not specified)	7,240 (35.2)	15,371 (40.4)	
Peripheral vascular disease	60 (0.29)	533 (1.4)	
Heart failure	87 (0.42)	1,395 (3.7)	
CVD-related death	4,377 (21.3)	4,797 (12.6)	
Non-CVD related death	1,821 (8.9)	9,340 (24.5)	

n: total number; %: percentage/proportion; Mean age for incident stroke and mean age for subsequent MACE outcome are reported with standard deviation.

There was a total of 28,750 (59.5%) subsequent MACE outcomes recorded. Appendix D.4.3 and Appendix D.4.4 shows the distribution of subsequent MACE and its constituent outcomes: recurrent stroke (n=19,896, 41.2%), cardiovascular mortality (n=4,797, 9.9%), and all-cause mortality (n=14,137, 29.3%). The overall incidence rate for subsequent MACE was 23.14 per 100 person-years (95% CI: 22.87–23.41), lower than the rate in the main analysis cohort (38.05; 95% CI: 37.71–38.39). Similar patterns for the incidence of subsequent major outcomes by sex and across the 5-year age groups, Appendix D.4.5, were obtained. The rate ratios by SES remained similar, Table 4.7. However, the incidence rates per 100 person-years were slightly lower when compared to the main analysis – recurrent stroke (15.43 vs 25.80), cardiovascular mortality (2.34 vs 3.44), and all-cause mortality (6.58 vs 7.37), Table 4.8.

Table 4.7 **Age- and sex-adjusted incidence rate ratio of subsequent major adverse outcomes, by socioeconomic status for patients with subsequent major adverse event after 30 days of index stroke** (n=48,306)

Socioeconomic status	MACE	Recurrent stroke	CVD-related mortality	All-cause mortality
1 (Highest SES)	Reference	Reference	Reference	Reference
2	1.06 (1.02 – 1.10)	1.06 (1.01 – 1.04)	1.06 (0.97 – 1.15)	1.03 (0.98 – 1.08)
3	1.10 (1.06 – 1.14)	1.08 (1.03 – 1.13)	1.07 (0.98 – 1.17)	1.04 (0.99 – 1.09)
4	1.13 (1.09 – 1.17)	1.08 (1.03 – 1.12)	1.15 (1.05 – 1.26)	1.12 (1.07 – 1.18)
5 (Lowest SES)	1.13 (1.09 – 1.18)	1.03 (0.99 – 1.08)	1.32 (1.21 – 1.45)	1.30 (1.23 – 1.37)

MACE: major adverse cardiovascular event

Table 4.8 **Incidence of subsequent major adverse outcomes for patients with subsequent major adverse events after 30 days of index stroke** (n=48,306)

	Follow-up time	Cases	Person-years*	Incidence rate (per 100 person-years)	Incidence rate ratio	*p*-value
MACE (All)	1.09 (0.31 – 2.91)	28,750	1,200	23.14 (22.87 – 23.41)		<0.0001
Men	1.09 (0.33 – 2.98)	12,973	575.95	22.53 (22.14 – 22.92)	Reference	
Women	1.09 (0.31 – 2.85)	15,777	666.62	23.67 923.30 – 24.04)	1.05 (1.03 – 1.08)	
Recurrent stroke (All)	1.00 (0.28 – 2.29)	19,896	1,300	15.43 (15.22 – 15.65)		<0.0001
Men	0.99 (0.28 – 2.32)	9,048	611.13	14.81 (14.50 – 15.11)	Reference	
Women	1.02 (0.29 – 2.28)	10,848	678.38	15.99 (15.69 – 16.30)	1.08 (1.05 – 1.11)	
Cardiovascular mortality (All)	1.33 (0.25 – 3.85)	4,797	2,100	2.34 (2.27 – 2.41)		<0.0001
Men	1.55 (0.33 – 4.29)	1,882	971.16	1.94 (1.85 – 2.03)	Reference	
Women	1.16 (0.23 – 3.62)	2,915	1,100	2.70 (2.60 – 2.80)	1.39 (1.31 – 1.48)	
All-cause mortality (All)	1.83 (0.44 – 4.66)	14,137	2,100	6.58 (6.47 – 6.69)		<0.0001
Men	1.90 (0.49 – 4.78)	5,836	1,000	5.78 (5.64 – 5.93)	Reference	
Women	1.83 (0.44 – 4.66)	8,301	1,100	7.28 (7.12 – 7.44)	1.26 (1.22 – 1.30)	
Coronary heart disease (All)	2.18 (0.82 – 4.63)	2,129	2,000	1.09 (1.05 – 1.14)		<0.0001
Men	2.17 (0.78 – 4.68)	1,149	921.94	1.25 (1.18 – 1.32)	Reference	
Women	2.18 (0.86 – 4.41)	980	1000	0.95 (0.89 – 1.01)	0.76 (0.70 – 0.83)	
Peripheral arterial disease (All)	1.88 (0.82 – 4.30)	533	2,000	0.27 (0.25 – 0.29)		<0.0001
Men	1.99 (0.81 – 4.21)	301	945.69	0.32 (0.28 – 0.36)	Reference	
Women	1.80 (0.84 – 4.38)	232	1,000	0.22 (0.19 – 0.25)	0.70 (0.58 – 0.83)	
Heart failure (All)	1.94 (0.69 – 4.68)	1,395	2,000	0.70 (0.67 – 0.74)		0.0001
Men	1.91 (0.70 – 4.52)	593	946.69	0.63 (0.58 – 0.68)	Reference	
Women	1.95 (0.69 – 4.90)	802	1,000	0.77 (0.72 – 0.82)	1.23 (1.10 – 1.37)	

* 100 person-years at risk; All – both men and women; Follow-up time – median follow-up time in years reported with interquartile range

4.5 Discussion

The incidence of stroke in the general population over the study period 1998-2017 was 109.20 per 100,000 person-years, with a peak incidence in 2004 and then a steady decline over the following years. The incidence of stroke was higher in women. Findings from this study show the greater burden of major adverse outcomes observed in women when compared to men after first non-fatal stroke is largely accounted for by age and socioeconomic status. The incidence of first stroke and subsequent all-cause mortality are higher in individuals in the lowest compared to the highest SES groups.

Findings from a meta-analysis using the Global Burden of Disease analytical technique (DisMod-MR), also estimated stroke incidence in the UK to be 120 strokes per 100,000 population.[212] A UK study with a smaller population of 1,657 individuals with acute vascular events in 9 Oxford primary care practices between 2002-2005, reported an incidence rate of 141 per 100,000 population (95% CI: 127-156) for ischaemic stroke with women having a higher incidence rate than men (147 vs 136 respectively).[205] By combining both primary care and hospital data in this study, out-of-hospital stroke events are more likely to be captured[213], hence more precise estimates of incidence.

The Quality and Outcome Framework (QOF), a pay-for-performance scheme covering a range of clinical and organisational areas in primary care, was introduced in the UK in April 2004.[89] Stroke was one of the 11 areas within the clinical domain. Although the QOF is a voluntary system, 99% of UK practices participate.[89] During the first year, the level of achievement exceeded the government anticipated level with an average of 83.4% of the allocated incentive payments claimed.[214] The rise in stroke incidence in 2004 could be attributed to the better recording due to the introduction of the QOF. This pay scheme also

incentivized primary prevention, in particular blood pressure and cholesterol control. These might have impacted the incidence of stroke post-2004.

In England, Wales, and Northern Ireland, a study by Wang et al., reported the average age for incident stroke to be 72 years for men and 78 for women[215] (71.4 and 76.8 years respectively in this study). Consistent with other studies, women had a higher rate of major adverse outcomes in the acute phase and were less likely to survive following stroke compared with males.[216,217] Women have more severe strokes than men[216,217] and the quality of care received by women with stroke is lower than that for men[218]. These are considered to be other possible reasons for the observed sex differences in outcome. From a public health perspective, is there is a need to monitor and compare stroke burden over time.[219] Population-based datasets such as electronic health records offer the opportunity to explore stroke burden and its association with risk factors such as socioeconomic status at a scale not previously possible.[220] Quantifying that association between socioeconomic status and stroke may help guide efforts aimed at reducing stroke burden, through local and focused secondary prevention interventions. This study is consistent with a number of studies indicating person-level measures of lowest SES is associated with a higher risk of first-ever stroke.[221,222] There is evidence of disparities in some aspects of stroke care and use of secondary prevention services for stroke – prescription of anticoagulation for atrial fibrillation, and timely admission to a specialist stroke unit.[222]

I was unable to identify any prior studies specifically describing the differences in incidence rates for subsequent major adverse outcomes after the first stroke. The closest recent studies assessing major adverse outcomes after the first stroke have been in selected populations with ischaemic stroke and/or above a pre-defined age.[206,207] Study by Sposato *et al*.[207], investigating sex-specific risks (not incident rate) of incident MACE in patients ≥66 years without known CVD comorbidities with first-ever ischaemic stroke and propensity-matched individuals

without stroke, found no sex difference in risk of incident MACE. Most women in their cohort were likely post-menopausal hence higher testosterone/oestradiol ratios or lower oestrogen levels in women may have evened the risk of MACE across sexes.[223] The higher rates of CHD and PVD in men compared to women in this study are in keeping with previous studies.[224,225] The risk of subsequent stroke within 90 days after acute TIA or minor stroke is high,[145,226] and this could explain the high proportion of TIA recorded in individuals with a subsequent event within 30 days of incident stroke. In a study by Bray *et al*., patients from most deprived areas had lower 1-year survival compared with those from less deprived areas, however, the effect of socioeconomic status was decreased after adjusting for baseline comorbidities.[222] Efforts to reduce disparities in stroke and subsequent outcomes need to address not only the access to good quality health care but also the social determinants of health and vascular risk factors earlier in life.

Index of Multiple Deprivation (IMD) 2015, a composite measure of relative area-level rather than individual-level deprivation, was used as a proxy measure of relative socioeconomic deprivation. Individual measures used to quantify individual-level deprivation/inequalities such as education level, income, employment type, and housing conditions, are not well reflected in area-level measures. With the same score being allocated to everyone within a given area, variations in deprivation within individuals within an area are not identified. This could explain why the impact of deprivation on stroke incidence and subsequent outcomes was not as marked as expected.

To my knowledge, this study is the most recent largest general population study to estimate the incidence of stroke and MACE following the first stroke using multiple data linkages to maximise ascertainment of stroke and MACE outcomes. The failure to use linked primary care and hospital data have been shown to lead to a substantial (25-50%) underestimation of the burden of cardiovascular disease like acute myocardial infarction.[209] CPRD, primary care database representative of

the UK general population,[87] is a rich source of longitudinal data and has been used to assess the incidence of a variety of health conditions.[87,227] I, therefore, assume the incidence of stroke and subsequent major adverse outcomes within the practices contributing data to CPRD accurately reflects the incidence in the wider UK population.

Considering the limitations of this study, case ascertainment is a potential limitation as the study is reliant on the presence of clinical codes indicative of stroke or any of the subsequent major adverse outcomes. Inaccurate recording may affect the estimates as is the case with all epidemiological studies using routine medical records. However, considering these conditions are QOF dependent and incentivised, the quality of the data remains high. It is not possible to be completely certain that subsequent coding of strokes does not relate to the ongoing care of the initial stroke hence the rates for stroke may be overestimated. Excluding patients without 12 months of data before incident stroke event minimises the likelihood of overestimating stroke incidence in this study. Due to the limited completeness of ethnicity information in people who were registered with CPRD up until 2006,[90] and small numbers within minority ethnic groups, ethnic differences were not assessed in this study.

4.6 Conclusion

This large population-based study linking national databases show there is significant morbidity and mortality within 2 years of the first stroke, with particular discrepancies across ages, sex, and socioeconomic status. Evidence of variation in major adverse cardiovascular event outcomes post-incident stroke by demographic and socioeconomic characteristics offers the opportunity to tailor secondary prevention and prioritise limited healthcare resources to those at greatest risk. This is discussed further in Chapter 8.

Summary

This chapter described the age, sex, and socioeconomic differences in the incidence of non-fatal stroke and subsequent MACE outcomes among adults. The next chapter will compare the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhage stroke and those with incident ischaemic stroke.

Chapter 5

Comparison of risk of serious cardiovascular events after haemorrhagic and ischaemic stroke

The previous chapter explored the age, sex, and socioeconomic differences in the rates of incident non-fatal stroke and first subsequent MACE outcomes using linked datasets. This chapter compares the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke.

A manuscript based on this study is under peer-review with the journal *Thrombosis and Haemostasis*:

Akyea, R. K., Georgiopoulos, G., Iyen, B., Kai, J., Qureshi, N., Ntaios, G. Comparison of risk of serious cardiovascular events after haemorrhagic versus ischaemic stroke: a population-based study.

5.1 Abstract

Background: Current guidelines recommend intensive preventive intervention to reduce subsequent very high cardiovascular risk in patients with ischaemic stroke. In contrast, there is no clear recommendation for patients with haemorrhagic stroke. This study compared the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and those with ischaemic stroke.

Methods and results: Patients aged ≥18 years with either incident haemorrhagic or ischaemic stroke between 1 January 1998 and 31 December 2017, and no prior history of the serious vascular event were identified from UK CPRD GOLD linked to HES APC data.

The study cohort included 32,091 patients with an overall follow-up of 381,237 person-years (median follow-up of 11.8 years). After adjusting for potential confounders, patients with incident haemorrhagic stroke had no significantly different risk of subsequent cardiovascular morbidity compared with patients with incident ischaemic stroke – CHD [HR: 0.86, 95% CI: 0.56 – 1.32], recurrent stroke [HR: 0.92, 95% CI: 0.83 – 1.02], PVD [HR: 1.15, 95% CI: 0.56 – 2.38], or heart failure [HR: 1.03, 95% CI: 0.61 – 1.74]. Patients with incident haemorrhagic stroke, however, had a significantly higher risk of subsequent CVD-related mortality [HR: 2.35, 95% CI: 2.04 – 2.72] and all-cause mortality [HR: 2.16, 95% CI: 1.94 – 2.41].

Propensity-score matched analysis of 1,039 patients with haemorrhagic stroke and 1,039 with ischaemic stroke showed similar risk in subsequent cardiovascular morbidity – CHD [stratified hazard ratio (sHR): 0.92, 95% CI: 0.55 – 1.54], recurrent stroke [sHR: 0.93, 95% CI: 0.82 – 1.02)], PVD [sHR: 1.04 95% CI: 0.45 – 2.41], or heart failure [HR: 0.71, 95% CI:0.39 – 1.27].

Conclusions: The risk of subsequent cardiovascular morbidity outcomes were similar between patients with incident haemorrhagic or ischaemic stroke. Patients with either incident haemorrhagic or ischaemic stroke should, therefore, be regarded as populations at very high risk of subsequent cardiovascular morbidity.

5.2 Introduction

Patients with ischemic stroke are considered a very-high risk population for subsequent cardiovascular events and current guidelines recommend intensive preventive strategies to reduce the cardiovascular risk.[228] In contrast, the amount of evidence about the overall cardiovascular risk in patients with haemorrhagic stroke is limited. Previous studies reported rates of cardiovascular events in patients with previous haemorrhagic stroke, but most of them were hospital-based analyses that were prone to selection bias and focused on selected outcomes over short follow-up.[229–232] A recent analysis of patients with previous haemorrhagic stroke in two population-based studies reported a rate of 7.9 serious vascular events per 100 patient-years.[232] Another analysis of four population-based studies estimated that the rate of arterial ischaemic events, ischaemic stroke and myocardial infarction is 2-3 times higher in patients with previous intracerebral haemorrhage compared to patients without.[233] To date, there is no reliable evidence to compare the risk of future cardiovascular events in patients with haemorrhagic and those with ischaemic stroke and therefore, it is unclear whether patients with haemorrhagic stroke should be regarded as a population with a very high risk of subsequent cardiovascular disease (CVD).

Using a large population-based cohort in the United Kingdom, I aimed to compare the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke after controlling for confounders or simulating inter-group differences in individual characteristics.

5.3 Methods

Data source

This prospective population-based cohort study used the UK CPRD GOLD[87] linked to HES APC,[208] national mortality data,[95] and social deprivation data.[97]

Study population

The study cohort of patients with the first record of non-fatal stroke in either CPRD GOLD or HES APC between 1 January 1998 and 31 December 2017 has been previously described in Chapter 2 (Section 2.3.1).

Cohort demographics and baseline characteristics

Age was defined at the time of the incident stroke. Ethnicity was categorised into six groups: Asian, Black, Mixed, Other, White and unknown.[90] Socioeconomic status based on the English IMD 2015,[97] described in Chapter 2 (Section 2.2.4), was categorised into quintiles (quintile 1 – least deprived group, to quintile 5 – most deprived group). Medication prescriptions (issue of a prescription) at baseline were defined as a prescription within 12 months before the incident stroke. For cholesterol (low-density lipoprotein (LDL), high-density lipoprotein (HDL) and total), body mass index (BMI), and blood pressure measures (diastolic and systolic), the most recent values/measures within 24 months before incident stroke were used. All other comorbidities were defined based on the latest record before the incident stroke.

Outcomes

First subsequent coronary heart disease (CHD), recurrent stroke, PVD and heart failure after incident stroke were the primary outcomes. Composite MACE, cardiovascular-related mortality, and all-cause mortality were considered secondary outcomes. MACE was defined as a composite of new-onset coronary heart disease, recurrent stroke, peripheral vascular disease, and heart failure.

Outcome events were based on records from the linked data sources (CPRD, HES APC, or ONS registry).

Statistical analysis

Continuous variables were summarised as mean (SD) or median (IQR); nominal variables were presented as counts and valid percentages. Normal distribution was graphically assessed by histograms and P-P plots. Kruskal-Wallis test for continuous data and the chi-squared test for categorical data were used to compare baseline characteristics. The level of missing values ranged between 19.4% for blood pressure measures to 69.9% for LDL-C. Details on the proportion of missingness are provided in Appendix E.5.1.

Complete-case analysis

The primary analysis was performed on the complete-case cohort and included two sub-analyses: one for the entire population of the complete-case cohort, and the other for a propensity-score matched population of the complete-case cohort. I used a multivariable probit regression model to calculate propensity scores for the conditional probability of classification (ischaemic versus haemorrhagic stroke) in 5,368 patients with ischaemic and 1,045 patients with haemorrhagic stroke. The propensity score (PS) matching model included age, sex, general practice, smoking status, socioeconomic status (IMD), blood pressure, BMI, HDL-C, LDL-C, diagnosis of atrial fibrillation, alcohol problem, cancer, dementia, diabetes mellitus, dyslipidaemia, hypertension, severe mental illness, transient ischaemic attack, family history of cardiovascular disease, a prescription of antihypertensive, anticoagulant, antidepressant, antiplatelet, diuretic, NSAIDs, opioids and potency of prescribed statin. I matched 2,078 patients with incident haemorrhagic and ischaemic stroke using a 1:1 greedy matching algorithm of nearest neighbour with a calliper of 0.01 and no replacement – Appendices E.5.2 – E.5.4.

Cox proportion hazards models were used to estimate the hazard ratio (HR) with a 95% confidence interval (95% CI) for subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke. Multivariable Cox models adjusting for pre-specified covariates based on relevant literature or biological plausibility [age at time of incident stroke, sex, socioeconomic status, smoking status, body mass index, blood pressure, cholesterols (high-density lipoprotein, low-density lipoprotein, and total), diagnosis of alcohol problem, atrial fibrillation, cancer, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, a prescription of antihypertensive, anticoagulant, antidiabetic, and potency of prescribed statin] were used for the entire cohort (non-PS-matched). For composite MACE outcomes, patients were censored at the time of the first outcome event. Cox regression models with shared frailty on matched sets were used for the PS-matched cohort, to account for the 'cluster effect' within matched pairs.[234] Kaplan-Meier curves were calculated to determine outcomes segregated by incident stroke sub-type (haemorrhagic vs. ischaemic). The log-rank test was used to compare the equality of the cumulative incidence plots between the stroke sub-type groups in the full cohort, while the stratified log-rank was used in the PS-matched cohort.[235]

Multiple imputation analyses in the overall cohort

I also performed a multiple imputation analysis in the overall cohort which included two sub-analyses: one at the entire population of the overall cohort, and one at a propensity-score matched population of the overall cohort. To estimate missing values for BMI, systolic and diastolic blood pressures, HDL-C, LDL-C and total cholesterol levels, multiple imputation by chained equations was used to generate imputed datasets as described in Chapter 2 (Section 2.3.4).[108–110] The imputed datasets were pooled into a single dataset using Rubin's rules.[111] The propensity score matching methodology was undertaken as previously described

– Appendices E.5.5 – E.5.7. These additional analyses were performed to evaluate the robustness of the findings due to potential bias from the use of imputed values for the analyses.

Landmark analysis

To minimise the potential impact of incident stroke severity on subsequent mortality during the early/subacute phase, further 3- and 6-months landmark analyses, as described in Chapter 2 (Section 2.4.4), were performed – patients with subsequent outcomes within the landmark periods were excluded.

All statistical analyses were performed using Stata SE version 17 (StataCorp LP). An alpha level of 0.05 was used for all analyses and all tests were 2-tailed. The study flow diagram is shown in Figure 5.1.

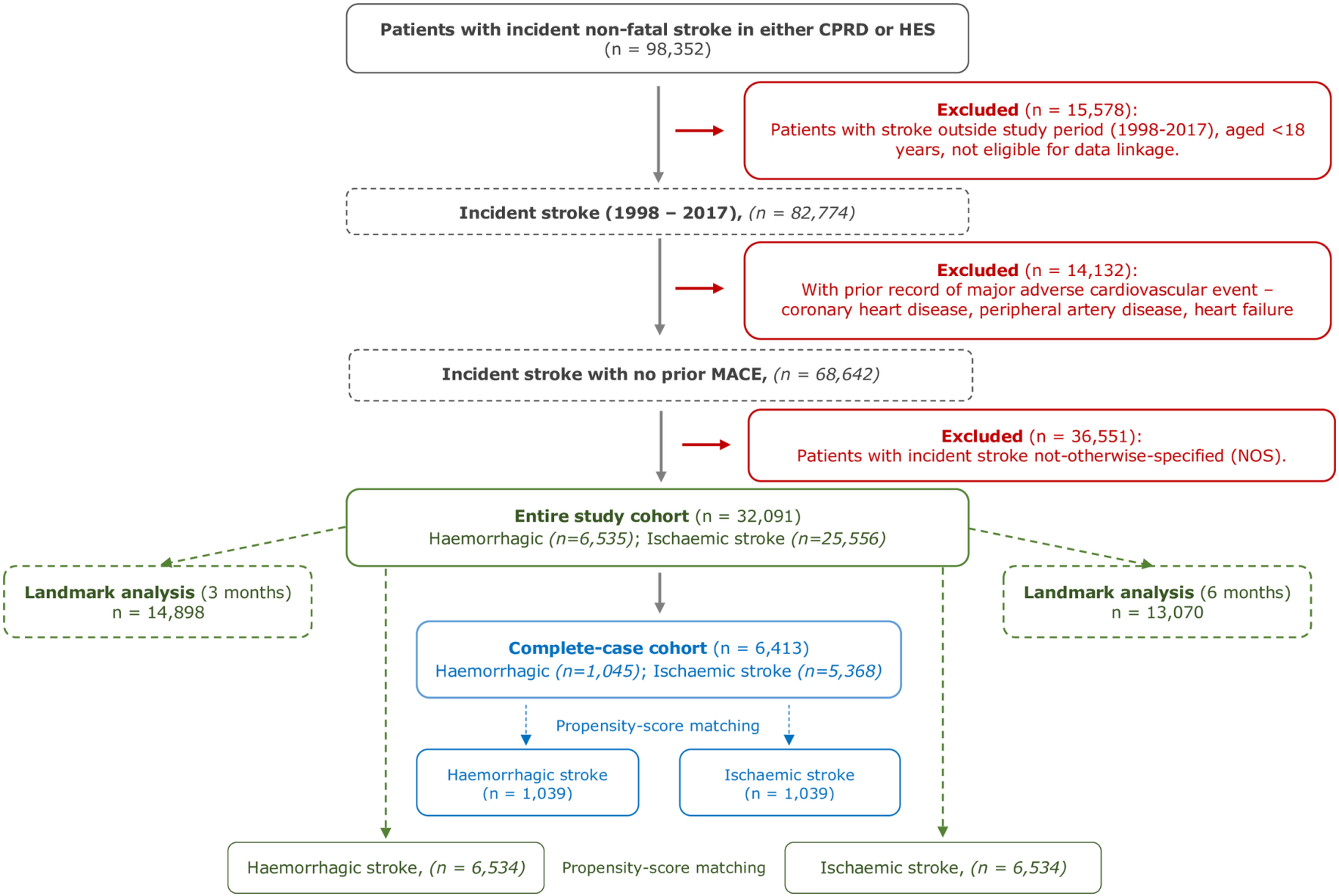


Figure 5.1 Study flow diagram

5.4 Results

Clinical characteristics

There were 32,091 patients who developed either incident haemorrhagic or ischaemic stroke events between 1 January 1998 and 31 December 2017 with 16,834 (52.5%) being women. Of these, 6,413 patients had complete data for all study variables – 1,045 (16.3%) had an incident haemorrhagic stroke and 5,368 (83.7%) had an incident ischaemic stroke event. The median age was 75 years. Patients with ischaemic stroke more often had diabetes mellitus or chronic kidney disease at the time of incident stroke event (Table 5.1). The overall follow-up for the cohort was 381,237.92 patient-years, corresponding to a median of 11.8 years (IQR: 6.9 – 16.2).

Table 5.1 **Characteristics of the study population with complete data at the time of incident stroke according to stroke sub-type** (n=6,413)

Characteristics	**Entire cohort** 6,413 (100%)	**Haemorrhagic** 1,045 (16.3%)	**Ischaemic** 5,368 (83.7%)	***p*-value**
Follow-up, median (IQR)	13.27 (8.37 – 17.68)	12.37 (7.65 – 16.99)	13.42 (8.56 – 17.80)	0.0002
Females	3,217 (50.2)	511 (48.9)	2,706 (50.4)	0.372
Age (years), mean (SD)	75 (66 – 81)	74 (66 – 81)	75 (66 – 82)	0.0758
Ethnicity				0.614
Asian	201 (3.1)	36 (3.4)	165 (3.1)	
Black	110 (1.7)	20 (1.9)	90 (1.7)	
Mixed	16 (0.3)	14 (0.3)	2 (0.2)	
Other	74 (1.2)	17 (1.6)	57 (1.1)	
White	5,821 (90.8)	937 (89.7)	4,884 (91.0)	
Unknown	191 (3.0)	33 (3.2)	158 (2.9)	
Socioeconomic status				0.009
1 (Least deprived)	1,348 (21.0)	262 (25.1)	1,086 (20.2)	
2	1,359 (21.2)	210 (20.1)	1,149 (21.4)	
3	1,352 (21.1)	221 (21.2)	1,131 (21.1)	
4	1,247 (19.4)	178 (17.0)	1,069 (19.9)	
5 (Most deprived)	1,100 (17.2)	172 (16.5)	928 (17.3)	
Unknown	7 (0.1)	2 (0.2)	5 (0.1)	
Current smokers	1,153 (18.0)	168 (16.1)	985 (18.4)	0.080
DBP (mmHg)	80 (70 – 85)	80 (71 – 85)	80 (70 – 84)	0.0483
SBP (mmHg)	140 (129 – 149)	140 (130 – 150)	139 (129 – 148)	0.0318
HDL cholesterol (mmol/L)	1.38 (1.10 – 1.70)	1.44 (1.20 – 1.80)	1.34 (1.10 – 1.68)	0.0001
LDL cholesterol (mmol/L)	2.80 (2.10 – 3.60)	2.70 (2.00 – 3.48)	2.80 (2.10 – 3.60)	0.0003
Total cholesterol (mmol/L)	4.90 (4.10 – 5.70)	4.80 (4.10 – 5.60)	4.90 (4.10 – 5.70)	0.0572
Alcohol problem	224 (3.5)	46 (4.4)	178 (3.3)	0.080
Atrial fibrillation	878 (13.7)	164 (15.7)	714 (13.3)	0.040
Cancer	1,248 (19.5)	230 (22.0)	1,018 (19.0)	0.023
Chronic kidney disease	1,332 (20.8)	190 (18.2)	1,142 (21.3)	0.024
Dementia	231 (3.6)	47 (4.5)	184 (3.4)	0.089
Diabetes mellitus	2,142 (33.4)	314 (30.1)	1,828 (34.1)	0.012
Type-1 diabetes	124 (1.9)	27 (2.6)	97 (1.8)	0.095
Type-2 diabetes	1,970 (30.7)	282 (27.0)	1,688 (31.5)	0.004
Dyslipidaemia	1,216 (19.0)	214 (20.5)	1,002 (18.7)	0.172
Family history of CVD	1,655 (25.8)	279 (26.7)	1,376 (25.6)	0.472
Hypertension	4,308 (67.2)	691 (66.1)	3,617 (67.4)	0.429
Severe mental illness	110 (1.7)	19 (1.8)	91 (1.7)	0.779
Transient ischaemic attack	512 (8.0)	71 (6.8)	441 (8.2)	0.121

Anti-coagulant	516 (8.1)	134 (12.8)	382 (7.1)	<0.001
Anti-diabetic	1,703 (26.6)	239 (22.9)	1,464 (27.3)	0.003
Anti-depressant	1,526 (23.8)	246 (23.5)	1,280 (23.9)	0.833
Anti-hypertensive	4,563 (71.2)	727 (69.6)	3,836 (71.5)	0.217
Anti-platelet	2,455 (38.3)	377 (36.1)	2,078 (38.7)	0.109
Diuretics	2,673 (41.7)	424 (40.6)	2,249 (41.9)	0.428
NSAIDS	1,715 (26.7)	261 (25.0)	1,454 (27.1)	0.158
Opioids	2,809 (43.8)	433 (41.4)	2,376 (44.3)	0.092
Statin				0.769
Low intensity	373 (5.8)	66 (6.3)	307 (5.7)	
Moderate intensity	2,208 (34.4)	348 (33.3)	1,860 (34.7)	
High intensity	619 (9.7)	100 (9.6)	519 (9.7)	

DBP: diastolic blood pressure; HDL: high-density lipoprotein; LDL: low-density lipoprotein; n: frequency/numbers; NSAIDs: non-steroidal anti-inflammatory drug; SBP: systolic blood pressure; SD: standard deviation; %: percent.

Complete-case analysis

Entire population

Of the 6,413 patients with incident stroke and complete data, 214 (3.3%) had a subsequent CHD outcome during follow-up [haemorrhagic: 24 (2.3%) versus ischaemic: 190 (3.5%)]; 3,140 (49.0%) had a recurrent stroke [haemorrhagic: 403 (38.6%) vs ischaemic: 2,737 (51.0%)]; 60 (0.9%) had PVD, and 134 (2.1%) had heart failure. After adjusting for potential confounders (Table 5.2), patients with incident haemorrhagic stroke had no significantly different risk of subsequent cardiovascular morbidity outcomes when compared with patients with incident ischaemic stroke – CHD [hazard ratio (HR), 0.86 (95% CI 0.56-1.32)]; recurrent stroke [HR, 0.92 (95% CI 0.83-1.02)], PVD [HR, 1.15 (95% CI 0.56-2.38)], or heart failure [HR, 1.03 (95% CI 0.61-1.74)].

Patients with incident haemorrhagic stroke had a significantly higher risk of subsequent CVD-related mortality [HR, 2.35 (95% CI 2.04-2.72)] and all-cause mortality [HR, 2.16 (95% CI 1.94-2.41)]. The cumulative incidence plots for the subsequent severe cardiovascular morbidity outcomes are presented in Figure 5.2.

Table 5.2 **Subsequent cardiovascular morbidity and mortality outcomes according to incident stroke sub-type for the entire and propensity-score matched complete case cohort**

Outcomes	Entire study cohort (n=6,413)				Propensity-score matched cohort (n=2,078)			
	Entire cohort 6,413 (100%)	**Ischaemic** 5,368 (83.7%)	**Haemorrhagic** 1,045 (16.3%)	***p*-value**	**Cohort** n=2,078 (100%)	**Ischaemic** n=1,039	**Haemorrhagic** n=1,039	***p*-value**
Coronary heart disease								
Number (percent)	214 (3.3)	190 (3.5)	24 (2.3)	0.041	60 (2.9)	36 (3.5)	24 (2.3)	0.116
Follow-up time	1.55 (0.22 – 3.79)	1.66 (0.22 – 3.79)	1.35 (0.39 – 3.81)	0.9386	1.62 (0.29 – 4.34)	2.17 (0.24 – 5.08)	1.35 (0.39 – 3.81)	0.4923
Incident rate [a]	1.18 (1.03 – 1.35)	1.19 (1.03 – 1.37)	1.07 (0.72 – 1.60)	-	1.10 (0.85 – 1.42)	1.11 (0.80 – 1.54)	1.08 (0.72 – 1.61)	-
Hazard ratio (95% CI)	–	Reference	0.86 (0.56 – 1.32)	0.490		Reference	0.92 (0.55 – 1.54)	0.752
Recurrent stroke								
Number (percent)	3,140 (49.0)	2,737 (51.0)	403 (38.6)	<0.001	927 (44.6)	526 (50.6)	401 (38.6)	<0.001
Follow-up time	0.06 (0.02 – 0.33)	0.06 (0.02 – 0.34)	0.05 (0.02 – 0.27)	0.1597	0.06 (0.02 – 0.30)	0.06 (0.02 – 0.36)	0.05 (0.02 – 0.25)	0.1840
Incident rate [a]	34.06 (32.89 – 35.28)	33.84 (32.60 – 35.14)	35.63 (32.32 – 39.29)	-	33.35 (31.27 – 35.57)	31.86 (29.25 – 34.70)	35.54 (32.22 – 39.19)	-
Hazard ratio (95% CI)	–	Reference	0.92 (0.83 – 1.02)	0.131		Reference	0.93 (0.82 – 1.06)	0.267
Peripheral vascular disease								
Number (percent)	60 (0.9)	51 (1.0)	9 (0.9)	0.785	22 (1.1)	13 (1.3)	9 (0.9)	0.391
Follow-up time	1.71 (0.85 – 3.79)	1.73 (0.81 – 3.75)	1.62 (1.16 – 4.47)	0.7094	1.61 (1.08 – 4.67)	1.51 (1.08 – 5.20)	1.62 (1.16 – 4.47)	0.9202
Incident rate [a]	0.32 (0.25 – 0.42)	0.31 (0.24 – 0.41)	0.40 (0.21 – 0.76)	-	0.40 (0.26 – 0.60)	0.39 (0.23 – 0.68)	0.40 (0.21 – 0.77)	-
Hazard ratio (95% CI)	–	Reference	1.15 (0.56 – 2.38)	0.705		Reference	1.04 (0.45 – 2.41)	0.932
Heart failure								
Number (percent)	134 (2.1)	117 (2.2)	17 (1.6)	0.253	51 (2.5)	34 (3.3)	17 (1.6)	0.016
Follow-up time	1.49 (0.41 – 3.41)	1.50 (0.41 – 3.28)	1.35 (0.60 – 3.41)	0.7131	1.17 (0.54 – 3.75)	1.14 (0.41 – 3.75)	1.35 (0.60 – 3.41)	0.5758
Incident rate [a]	0.73 (0.61 – 0.86)	0.72 (0.60 – 0.87)	0.75 (0.47 – 1.21)	-	0.92 (0.70 – 1.21)	1.04 (0.74 – 1.45)	0.76 (0.47 – 1.22)	-
Hazard ratio (95% CI)	–	Reference	1.03 (0.61 – 1.74)	0.898		Reference	0.71 (0.39 – 1.27)	0.249

Outcomes	Entire study cohort (n=6,413)				Propensity-score matched cohort (n=2,078)			
	Entire cohort 6,413 (100%)	**Ischaemic** 5,368 (83.7%)	**Haemorrhagic** 1,045 (16.3%)	***p*-value**	**Cohort** n=2,078 (100%)	**Ischaemic** n=1,039	**Haemorrhagic** n=1,039	***p*-value**
Major adverse cardiovascular event (composite)								
Number (percent)	3,548 (55.3)	3,095 (57.7)	453 (43.4)	<0.001	1,060 (51.0)	609 (58.6)	451 (43.4)	0.213
Follow-up time	0.7 (0.02 – 0.68)	0.08 (0.02 – 0.72)	0.07 (0.02 – 0.45)	0.1861	0.07 (0.02 – 0.65)	0.08 (0.02 – 0.86)	0.07 (0.02 – 0.45)	0.0794
Incident rate [a]	41.97 (40.61 – 43.37)	41.80 (40.35 – 43.29)	43.16 (39.37 – 47.33)	-	41.34 (38.92 – 43.90)	40.14 (37.07 – 43.46)	43.07 (39.28 – 47.24)	-
Hazard ratio (95% CI)	–	Reference	0.93 (0.84 – 1.02)	0.130		Reference	0.92 (0.82 – 1.03)	0.166
Cardiovascular-related mortality								
Number (percent)	993 (15.5)	726 (13.5)	267 (25.6)	<0.001	398 (19.2)	133 (12.8)	265 (25.5)	<0.001
Follow-up time	0.05 (0.01 – 0.35)	0.07 (0.02 – 0.67)	0.02 (0.01 – 0.07)	0.0001	0.02 (0.01 – 0.16)	0.08 (0.01 – 0.54)	0.02 (0.01 – 0.07)	0.0001
Incident rate [a]	5.23 (4.92 – 5.57)	4.36 (4.06 – 4.69)	11.43 (10.14 – 12.89)	-	6.99 (6.34 – 7.71)	3.95 (3.33 – 4.68)	11.40 (10.11 – 12.86)	-
Hazard ratio (95% CI)	–	Reference	2.35 (2.04 – 2.72)	<0.001		Reference	2.36 (1.93 – 2.90)	<0.001
All-cause mortality								
Number (percent)	1,786 (27.9)	1,346 (25.1)	440 (42.1)	<0.001	680 (32.7)	243 (23.4)	437 (42.1)	<0.001
Follow-up time	0.18 (0.02 – 2.25)	0.28 (0.04 – 2.88)	0.05 (0.01 – 0.70)	0.0001	0.10 (0.01 – 1.43)	0.29 (0.02 – 2.81)	0.05 (0.01 – 0.72)	0.0001
Incident rate [a]	9.09 (8.68 – 9.52)	7.83 (7.42 – 8.26)	17.93 (16.33 – 19.68)	-	11.48 (10.65 – 12.38)	6.99 (6.16 – 7.92)	17.88 (16.28 – 19.64)	-
Hazard ratio (95% CI)	–	Reference	2.16 (1.94 – 2.41)	<0.001		Reference	2.24 (1.92 – 2.62)	<0.001

Follow-up time: Time from incident stroke event to mortality outcome reported as median with interquartile range. CI, confidence interval; HR, hazard ratio

[a] Incident rate per 100 person-years.

Model adjusted for age at the time of incident stroke, sex, socioeconomic status, smoking status, body mass index, blood pressure, cholesterols (high-density lipoprotein, low-density lipoprotein, and total), diagnosis of an alcohol problem, atrial fibrillation, cancer, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, a prescription of antihypertensive, anticoagulant, antidiabetic, and potency of prescribed statin.

Stratified hazard ratio (that is, Cox regression models with shared frailty) reported for propensity-score matched cohort.

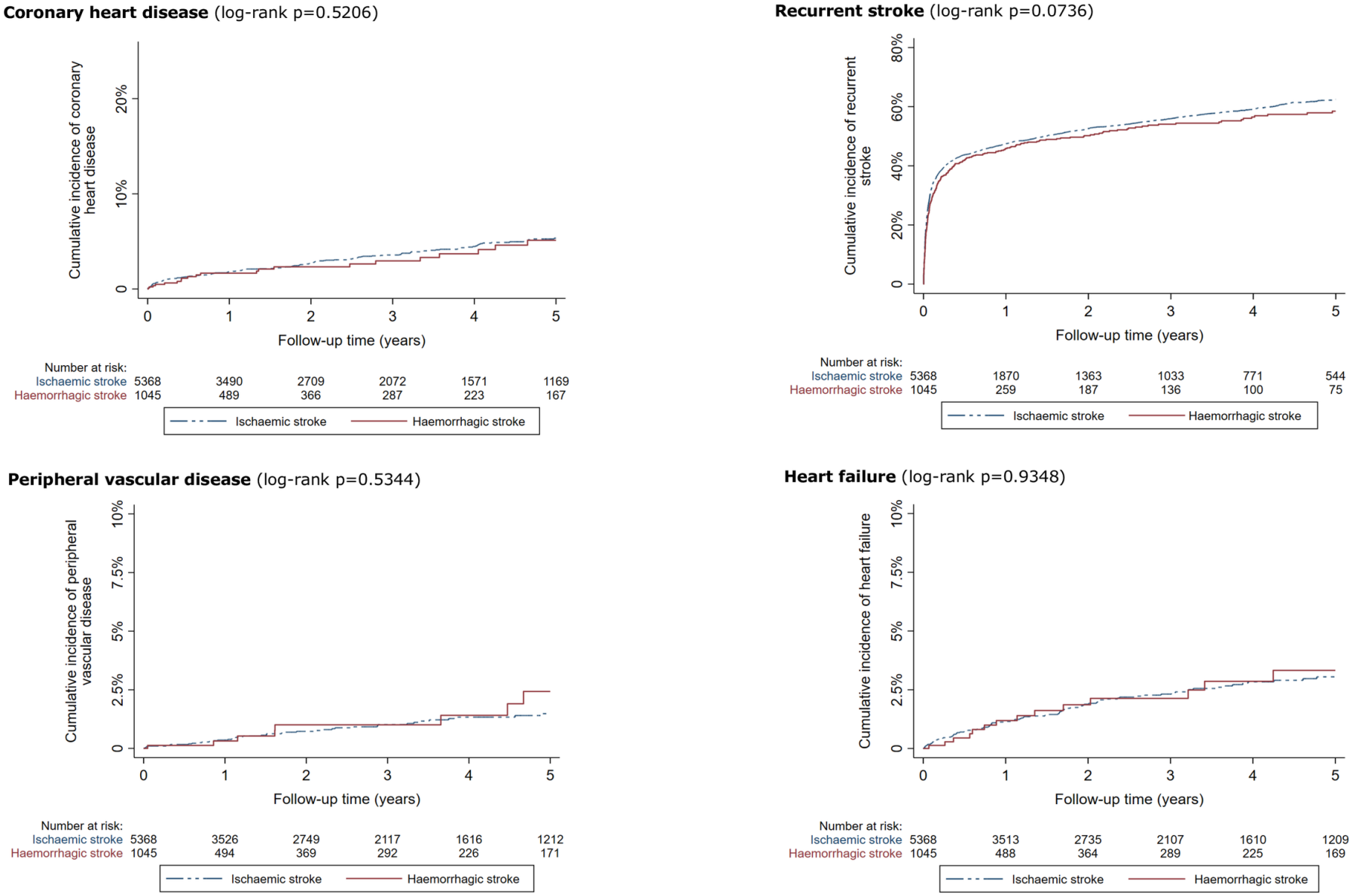
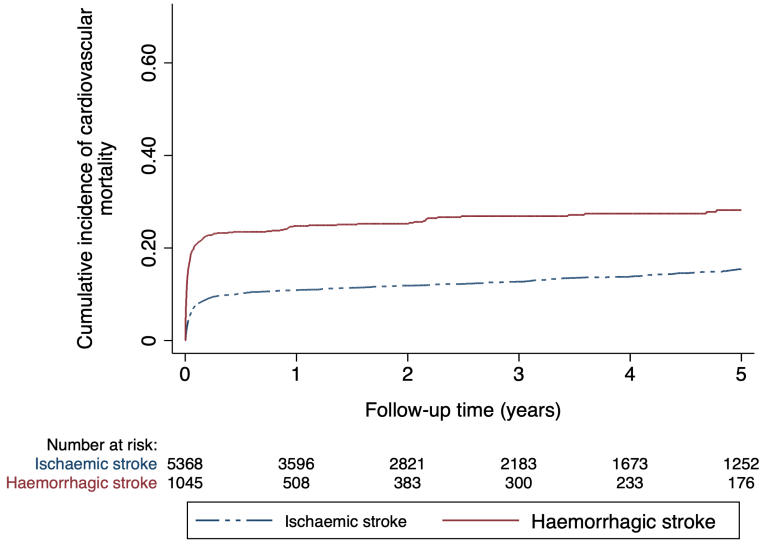
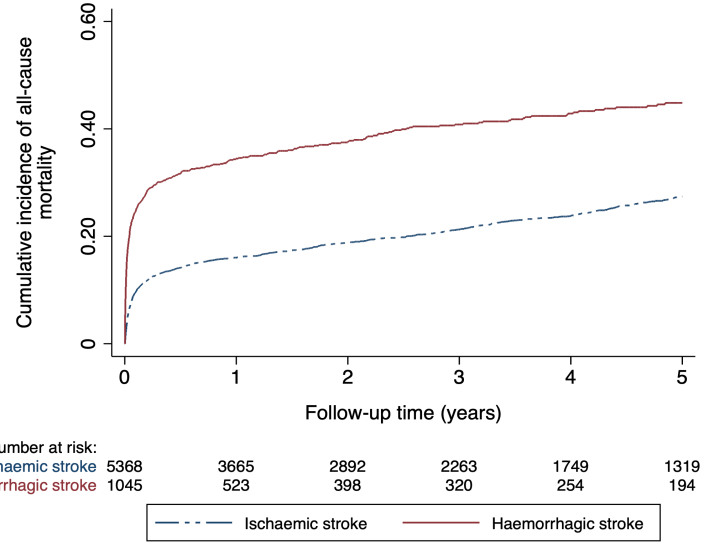


Figure 5.2 **Cumulative incidence plot for subsequent severe cardiovascular morbidity outcomes** (entire complete case cohort, n=6,413)

Cardiovascular-related mortality (log-rank p<0.0001)



All-cause mortality (log-rank p<0.0001)



Composite MACE (log-rank p=0.0665)

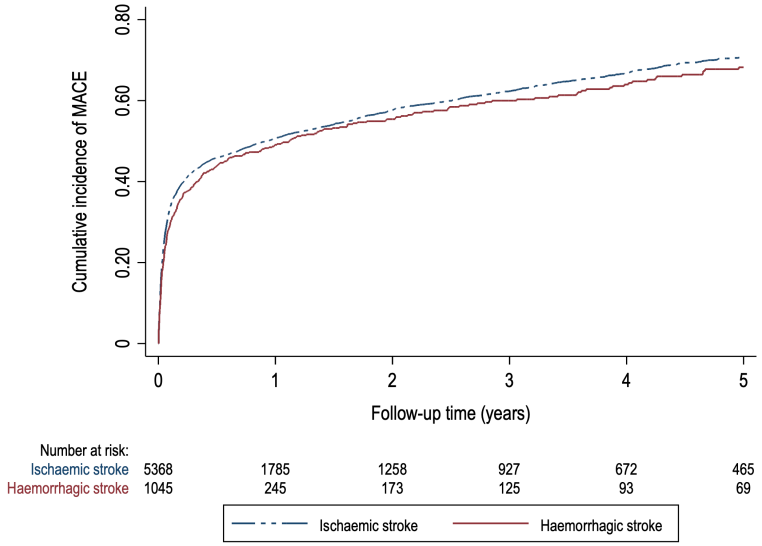


Figure 5.2 **Cumulative incidence plot for subsequent severe cardiovascular morbidity outcomes** (entire complete case cohort, n=6,413)

Propensity-score matched analysis

For the propensity-score matched analysis of the complete-case population, 1,039 patients with haemorrhagic stroke were matched with 1,039 with ischaemic stroke. Risk of subsequent cardiovascular morbidity outcomes were not significantly different between patients with incident haemorrhagic stroke compared with those with incident ischaemic stroke – CHD [stratified hazard ratio (sHR), 0.92 (95% CI 0.55-1.54)]; recurrent stroke [sHR, 0.93 (95% CI 0.82-1.06)], PVD [sHR, 1.04 (95% CI 0.45-2.41)], or heart failure [HR, 0.71 (95% CI 0.39-1.27)].

The risk of subsequent mortality outcomes was significantly higher in patients with incident haemorrhagic stroke – cardiovascular-related mortality [sHR, 2.36 (95% CI 1.93-2.90)] and all-cause mortality [sHR, 2.24 (95% CI 1.92-2.62)]– Table 5.2.

Multiple imputation analyses in the overall cohort

Entire population

In this analysis, the overall study cohort of 32,091 patients with incident stroke events and with missing values imputed was used – 6,535 (20.5%) of these patients had an incident haemorrhagic stroke and 25,556 (79.6%) had an incident ischaemic stroke event. The characteristic of the overall cohort is presented in Table 5.3. After adjusting for potential confounders in the entire cohort, patients with incident haemorrhagic as compared with those with ischaemic stroke had lower risk of subsequent CHD [n=926 (2.9%), HR 0.67 (95% CI 0.55-0.82)] and an increased risk of subsequent CVD-related mortality [n=6,001 (18.7%), HR 2.26 (95% CI 2.13-2.39)], and all-cause mortality [n=10,675 (33.3%), HR 1.95 (95% CI 1.86-2.03)]. The cumulative incidence plots for cardiovascular morbidity and mortality outcomes are presented in Figure 5.3.

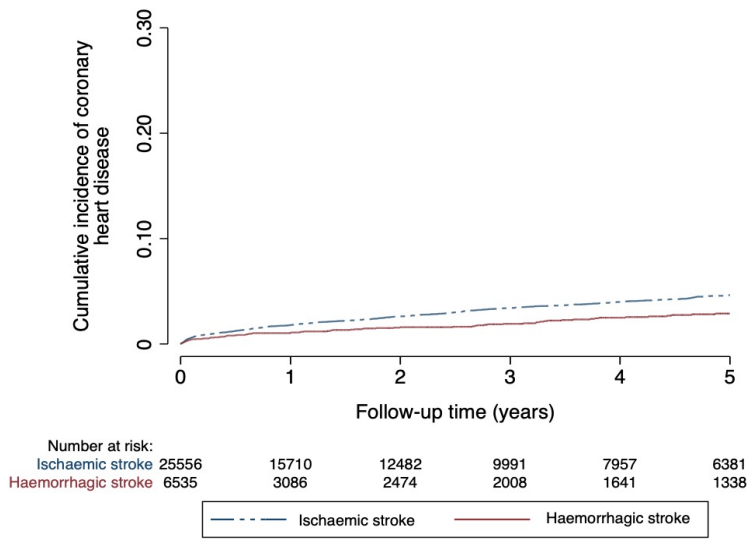
Table 5.3 Characteristics of the entire study population at the time of incident stroke according to stroke subtype (n=32,091)

Characteristics	Entire cohort 32,091 (100%)	Haemorrhagic 6,535 (20.4%)	Ischaemic 25,556 (79.6%)	*p*-value
Follow-up, median (IQR)	12.1 (7.0 – 16.6)	10.9 (6.0 – 15.8)	12.4 (7.4 – 16.7)	0.0001
Females	16,834 (52.5)	3,375 (51.6)	13,459 (52.7)	0.141
Age (years), mean (SD)	75 (64 – 83)	73 (61 – 82)	76 (65 – 84)	0.0001
Ethnicity				<0.001
Asian	482 (1.5)	112 (1.7)	370 (1.5)	
Black	297 (0.9)	85 (1.3)	212 (0.8)	
Mixed	64 (0.2)	18 (0.3)	46 (0.2)	
Other	245 (0.8)	59 (0.9)	186 (0.7)	
White	28,981 (90.3)	5,738 (87.8)	23,243 (91.0)	
Unknown	2,022 (6.3)	523 (8.0)	1,499 (5.9)	
Socioeconomic status				0.003
1 (Least deprived)	6,937 (21.6)	1,513 (23.2)	5,424 (21.2)	
2	7,072 (22.0)	1,450 (22.2)	5,622 (22.0)	
3	6,901 (21.5)	1,389 (21.3)	5,512 (21.6)	
4	5,960 (18.6)	1,183 (18.1)	4,777 (18.7)	
5 (Most deprived)	5,172 (16.1)	986 (15.1)	4,186 (16.4)	
Unknown	49 (0.2)	14 (0.2)	35 (0.1)	
Current smokers	6,113 (19.1)	1,132 (17.3)	4,981 (19.5)	<0.001
DBP (mmHg)	80 (74 – 84)	80 (76 – 84)	80 (74 – 84)	0.0001
SBP (mmHg)	140 (130 – 148)	140 (132 – 148)	140 (130 – 148)	0.0465
HDL cholesterol (mmol/L)	1.47 (1.30 – 1.63)	1.49 (1.35 – 1.65)	1.46 (1.30 – 1.62)	0.0001
LDL cholesterol (mmol/L)	2.97 (2.70 – 3.25)	2.97 (2.70 – 3.21)	2.97 (2.69 – 3.27)	0.0310
Total cholesterol (mmol/L)	5.10 (4.75 – 5.44)	5.10 (4.80 – 5.40)	5.10 (4.73 – 5.45)	0.9201
Alcohol problem	1,017 (3.2)	261 (4.0)	756 (3.0)	<0.001
Atrial fibrillation	3,455 (10.8)	585 (9.0)	2,870 (11.2)	<0.001
Cancer	5,359 (16.7)	1,114 (17.1)	4,245 (16.6)	0.399
Chronic kidney disease	3,931 (12.3)	616 (9.4)	3,315 (13.0)	<0.001
Dementia	1,245 (3.9)	287 (4.4)	958 (3.8)	0.016
Diabetes mellitus	3,910 (12.2)	576 (8.8)	3,334 (13.1)	<0.001
Type-1 diabetes	280 (0.9)	47 (0.7)	233 (0.9)	0.135
Type-2 diabetes	3,361 (10.5)	466 (7.1)	2,895 (11.3)	<0.001
Dyslipidaemia	2,840 (8.9)	514 (7.9)	2,326 (9.1)	0.002

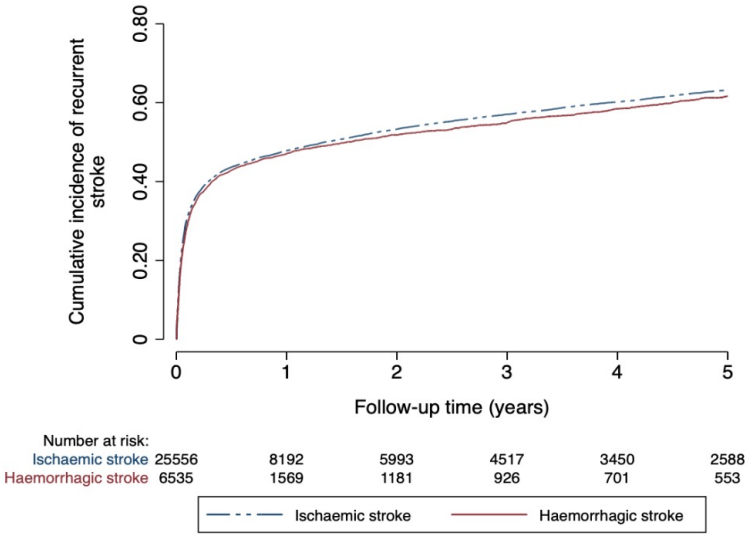
Characteristics	Entire cohort 32,091 (100%)	Haemorrhagic 6,535 (20.4%)	Ischaemic 25,556 (79.6%)	*p*-value
Family history of CVD	5,659 (17.6)	1,139 (17.4)	4,520 (17.7)	0.626
Hypertension	15,072 (47.0)	2,721 (41.6)	12,351 (48.3)	<0.001
Severe mental illness	399 (1.2)	81 (1.2)	318 (1.2)	0.975
Transient ischaemic attack	2,006 (6.3)	339 (5.2)	1,667 (6.5)	<0.001
Anti-coagulant	2,023 (6.3)	544 (8.3)	1,479 (5.8)	<0.001
Anti-diabetic	3,116 (9.7)	443 (6.8)	2,673 (10.5)	<0.001
Anti-depressant	6,757 (21.1)	1,353 (20.7)	5,404 (21.2)	0.434
Anti-hypertensive	15,700 (48.9)	2,729 (41.8)	12,971 (50.8)	<0.001
Anti-platelet	8,804 (27.4)	1,448 (22.2)	7,356 (28.8)	<0.001
Diuretics	10,884 (33.9)	1,882 (28.8)	9,002 (35.2)	<0.001
NSAIDS	7,971 (24.8)	1,546 (23.7)	6,425 (25.1)	0.013
Opioids	12,369 (38.5)	2,365 (36.2)	10,004 (39.2)	<0.001
Statin				<0.001
Low intensity	918 (2.9)	165 (2.5)	753 (3.0)	
Moderate intensity	4,914 (15.3)	799 (12.2)	4,115 (16.1)	
High intensity	1,206 (3.8)	205 (3.1)	1,001 (3.9)	

DBP: diastolic blood pressure; HDL: high-density lipoprotein; LDL: low-density lipoprotein; n: frequency/numbers; NSAIDs: non-steroidal anti-inflammatory drug; SBP: systolic blood pressure; SD: standard deviation; %: percent.

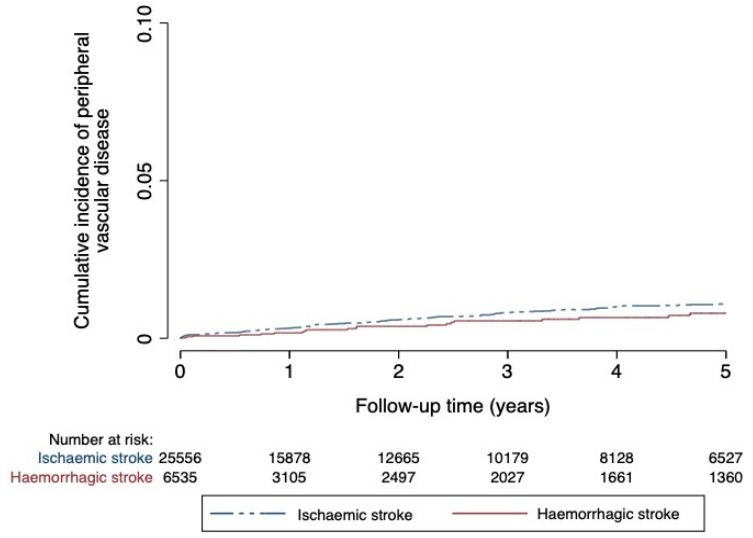
Coronary heart disease (log rank p<0.0001)



Recurrent stroke (log rank p=0.0146)



Peripheral vascular disease (log rank p=0.0969)



Heart failure (log rank p=0.0005)

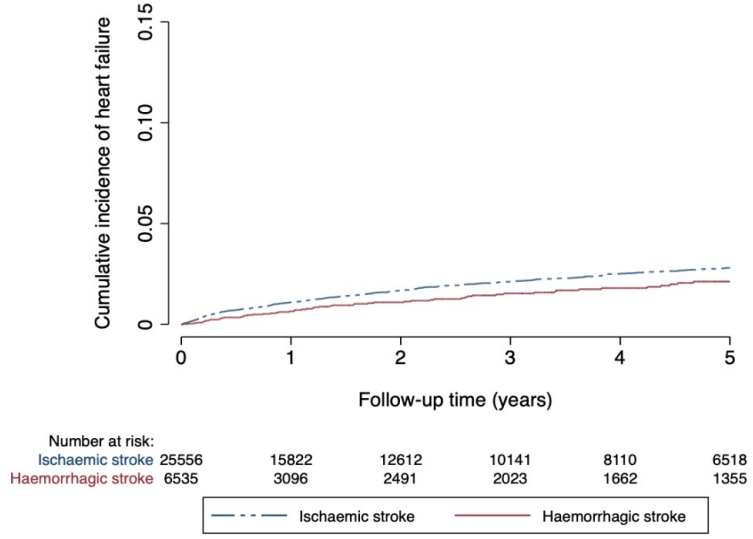
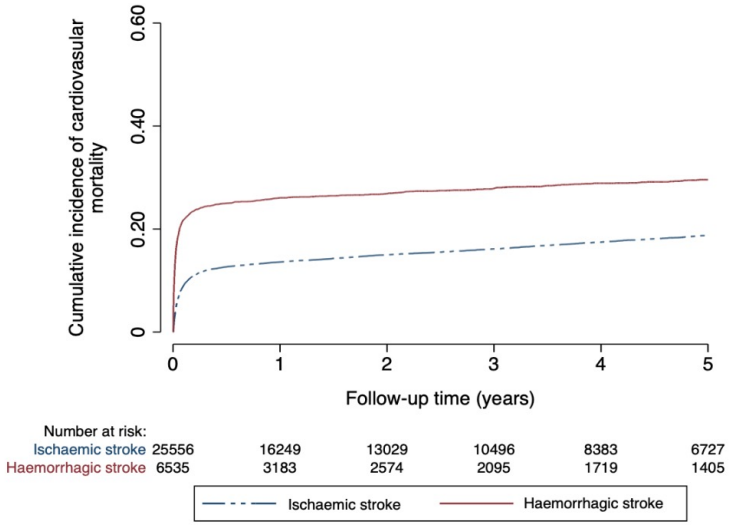
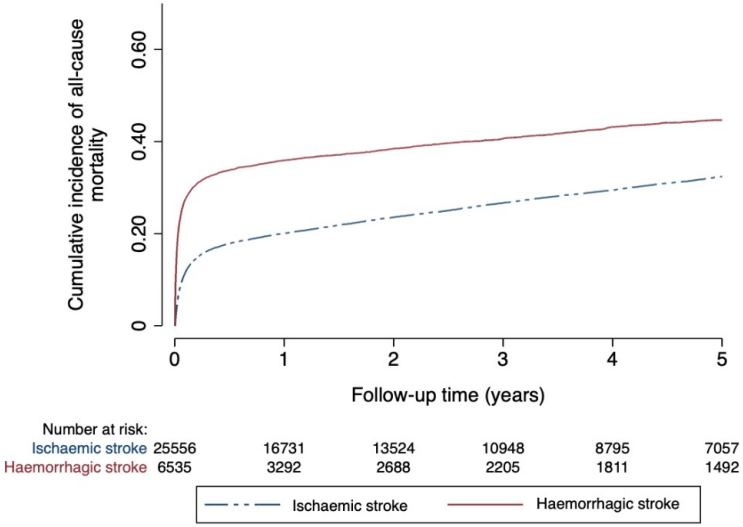


Figure 5.3 Cumulative incidence plot for subsequent morbidity and mortality outcomes for the entire cohort (n=32,091)

Cardiovascular-related mortality (log rank p<0.0001)



All-cause mortality (log rank p<0.0001)



Composite MACE (log rank p<0.0001)

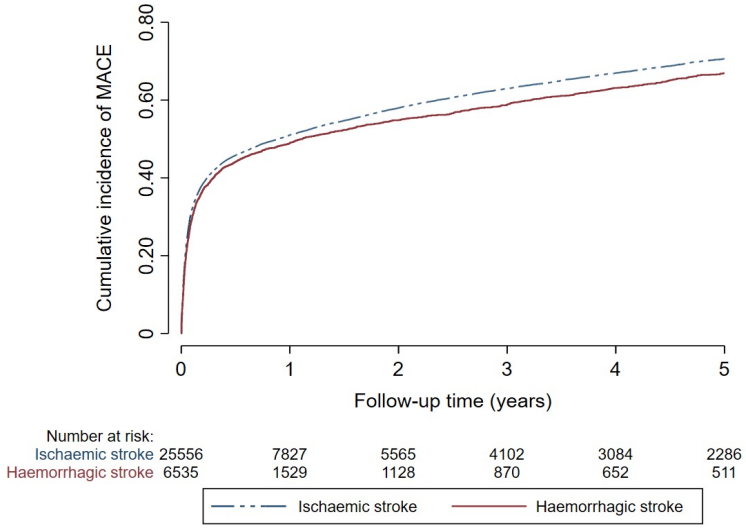


Figure 5.3 Cumulative incidence plot for subsequent morbidity and mortality outcomes for the entire cohort (n=32,091)

Propensity-score matched population

For the propensity-score matched analysis, 6,534 patients with haemorrhagic stroke were matched with 6,534 patients with ischaemic stroke. The findings were generally similar to findings from analysis using the entire study cohort of 32,091 patients (6,535 patients with incident haemorrhagic stroke and 25,556 with incident ischaemic stroke). The risk of subsequent CHD remained lower, and mortality (both CVD-related and all-cause) outcomes remained significantly higher in patients with incident haemorrhagic stroke compared with those with incident ischaemic stroke – Table 5.4. The cumulative incidence plots for cardiovascular morbidity and mortality outcomes are presented in Figure 5.4.

Table 5.4 Subsequent cardiovascular morbidity and mortality outcomes according to incident stroke sub-type for the entire and propensity-score matched cohort with imputed values

Outcomes	Entire study cohort (n=32,091)				Propensity-score matched cohort (n=13,068)			
	Entire cohort 32,091 (100%)	**Ischaemic** 25,556 (79.6%)	**Haemorrhagic** 6,535 (20.4%)	***p*-value**	**Cohort** n=13,068 (100%)	**Ischaemic** n=6,534	**Haemorrhagic** n=6,534	***p*-value**
Coronary heart disease								
Number (percent)	926 (2.9)	822 (3.2)	104 (1.6)	<0.001	328 (2.5)	224 (3.4)	104 (1.6)	<0.001
Follow-up time	1.34 (0.25 – 3.81)	1.33 (0.25 – 3.81)	1.38 (0.23 – 3.93)	0.9280	1.43 (0.19 – 4.45)	1.51 (0.17 – 4.60)	1.38 (0.23 – 3.93)	0.9674
Incident rate [a]	0.93 (0.88 – 1.00)	1.00 (0.94 – 1.07)	0.61 (0.50 – 0.74)	-	0.83 (0.74 – 0.92)	1.00 (0.88 – 1.14)	0.61 (0.50 – 0.74)	-
Hazard ratio (95% CI)	–	Reference	0.67 (0.55 – 0.82)	<0.001	-	Reference	0.60 (0.47 – 0.75)	<0.001
Recurrent stroke								
Number (percent)	15,417 (48.0)	12,818 (50.2)	2,599 (39.8)	<0.001	5,908 (45.2)	3,309 (50.6)	2,599 (39.8)	<0.001
Follow-up time	0.07 (0.02 – 0.38)	0.07 (0.02 – 0.39)	0.07 (0.02 – 0.32)	0.0106	0.07 (0.02 – 0.40)	0.07 (0.02 – 0.50)	0.07 (0.02 – 0.32)	0.0019
Incident rate [a]	33.41 (32.88 – 33.94)	33.59 (33.02 – 34.18)	32.51 (31.29 – 33.79)	-	32.21 (31.40 – 33.04)	31.97 (30.90 – 33.08)	32.52 (31.30 – 33.80)	-
Hazard ratio (95% CI)	–	Reference	0.93 (0.90 – 0.98)	0.002	-	Reference	0.96 (0.91 – 1.01)	0.115
Peripheral vascular disease								
Number (percent)	210 (0.7)	183 (0.7)	27 (0.4)	0.007	72 (0.6)	45 (0.7)	27 (0.4)	0.033
Follow-up time	1.71 (0.59 – 3.75)	1.69 (0.57 – 3.51)	2.26 (0.86 – 4.67)	0.3759	1.77 (0.60 – 3.38)	1.73 (0.33 – 2.58)	2.26 (0.86 – 4.67)	0.2090
Incident rate [a]	0.21 (0.18 – 0.24)	0.22 (0.19 – 0.25)	0.16 (0.11 – 0.23)	-	0.18 (0.14 – 0.23)	0.20 (0.15 – 0.26)	0.16 (0.11 – 0.23)	-
Hazard ratio (95% CI)	–	Reference	0.83 (0.55 – 1.25)	0.369	-	Reference	0.78 (0.48 – 1.26)	0.305
Heart failure								
Number (percent)	584 (1.8)	516 (2.0)	68 (1.0)	<0.001	179 (1.4)	111 (1.7)	68 (1.0)	0.001
Follow-up time	1.50 (0.39 – 4.08)	1.49 (0.37 – 4.12)	1.57 (0.58 – 3.72)	0.5862	1.66 (0.42 – 4.52)	1.70 (0.37 – 5.06)	1.57 (0.58 – 3.72)	0.8468
Incident rate [a]	0.58 (0.53 -0.63)	0.62 (0.57 – 0.67)	0.39 (0.31 – 0.50)	-	0.45 (0.39 – 0.52)	0.49 (0.40 – 0.59)	0.39 (0.31 – 0.50)	-
Hazard ratio (95% CI)	–	Reference	0.81 (0.62 – 1.04)	0.098	-	Reference	0.80 (0.59 – 1.09)	0.157

Outcomes	Entire study cohort (n=32,091)				Propensity-score matched cohort (n=13,068)			
	Entire cohort 32,091 (100%)	**Ischaemic** 25,556 (79.6%)	**Haemorrhagic** 6,535 (20.4%)	***p*-value**	**Cohort** n=13,068 (100%)	**Ischaemic** n=6,534	**Haemorrhagic** n=6,534	***p*-value**
Major adverse cardiovascular event (composite)								
Number (percent)	17,137 (53.4)	14,339 (56.1)	2,798 (42.8)	<0.001	6,487 (49.6)	3,689 (56.5)	2,798 (42.8)	<0.001
Follow-up time	0.08 (0.02 – 0.67)	0.08 (0.02 – 0.70)	0.07 (0.02 – 0.48)	0.0001	0.08 (0.02 – 0.66)	0.09 (0.02 – 0.84)	0.07 (0.02 – 0.48)	0.0001
Incident rate [a]	40.29 (39.69 – 40.90)	40.96 (40.30 – 41.64)	37.19 (35.83 – 38.59)	-	37.96 (37.05 – 38.90)	38.57 (37.34 – 39.83)	37.20 (35.84 – 38.60)	-
Hazard ratio (95% CI)	–	Reference	0.91 (0.87 – 0.95)	<0.001	-	Reference	0.92 (0.88 – 0.97)	<0.001
Cardiovascular-related mortality								
Number (percent)	6,001 (18.7)	4,248 (16.6)	1,753 (26.8)	<0.001	2,731 (20.9)	978 (15.0)	1,753 (26.8)	<0.001
Follow-up time	0.05 (0.01 – 0.34)	0.08 (0.02 – 0.69)	0.02 (0.01 – 0.07)	0.0001	0.03 (0.01 – 0.16)	0.08 (0.02 – 0.76)	0.02 (0.01 – 0.07)	0.0001
Incident rate [a]	5.79 (5.65 – 5.94)	4.95 (4.81 – 5.10)	9.84 (9.39 -10.31)	-	6.64 (6.40 – 6.70)	4.20 (3.94 – 4.47)	9.84 (9.39 – 10.31)	-
Hazard ratio (95% CI)	–	Reference	2.26 (2.13 – 2.39)	<0.001	-	Reference	2.12 (1.96 – 2.28)	<0.001
All-cause mortality								
Number (percent)	10,675 (33.3)	7,851 (30.7)	2,824 (43.2)	<0.001	4,673 (35.8)	1,849 (28.3)	2,824 (43.2)	<0.001
Follow-up time	0.15 (0.02 – 2.02)	0.24 (0.04 – 2.54)	0.04 (0.01 – 0.41)	0.0001	0.08 (0.01 – 1.27)	0.25 (0.04 – 2.50)	0.04 (0.01 – 0.41)	0.0001
Incident rate [a]	9.91 (9.72 – 10.10)	8.81 (8.62 – 9.01)	15.19 (14.64 – 15.76)	-	10.90 (10.59 – 11.22)	7.62 (7.28 – 7.97)	15.19 (14.64 – 15.76)	-
Hazard ratio (95% CI)	–	Reference	1.95 (1.86 – 2.03)	<0.001	-	Reference	1.85 (1.75 – 1.96)	<0.001

Follow-up time: Time from incident stroke event to mortality outcome reported as median with interquartile range. CI, confidence interval; HR, hazard ratio

[a] Incident rate per 100 person-years.

Model adjusted for age at the time of incident stroke, sex, socioeconomic status, smoking status, body mass index, blood pressure, cholesterols (high-density lipoprotein, low-density lipoprotein, and total), diagnosis of an alcohol problem, atrial fibrillation, cancer, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, a prescription of antihypertensive, anticoagulant, antidiabetic, and potency of prescribed statin.

Stratified hazard ratio (that is, Cox regression models with shared frailty) reported for propensity-score matched cohort.

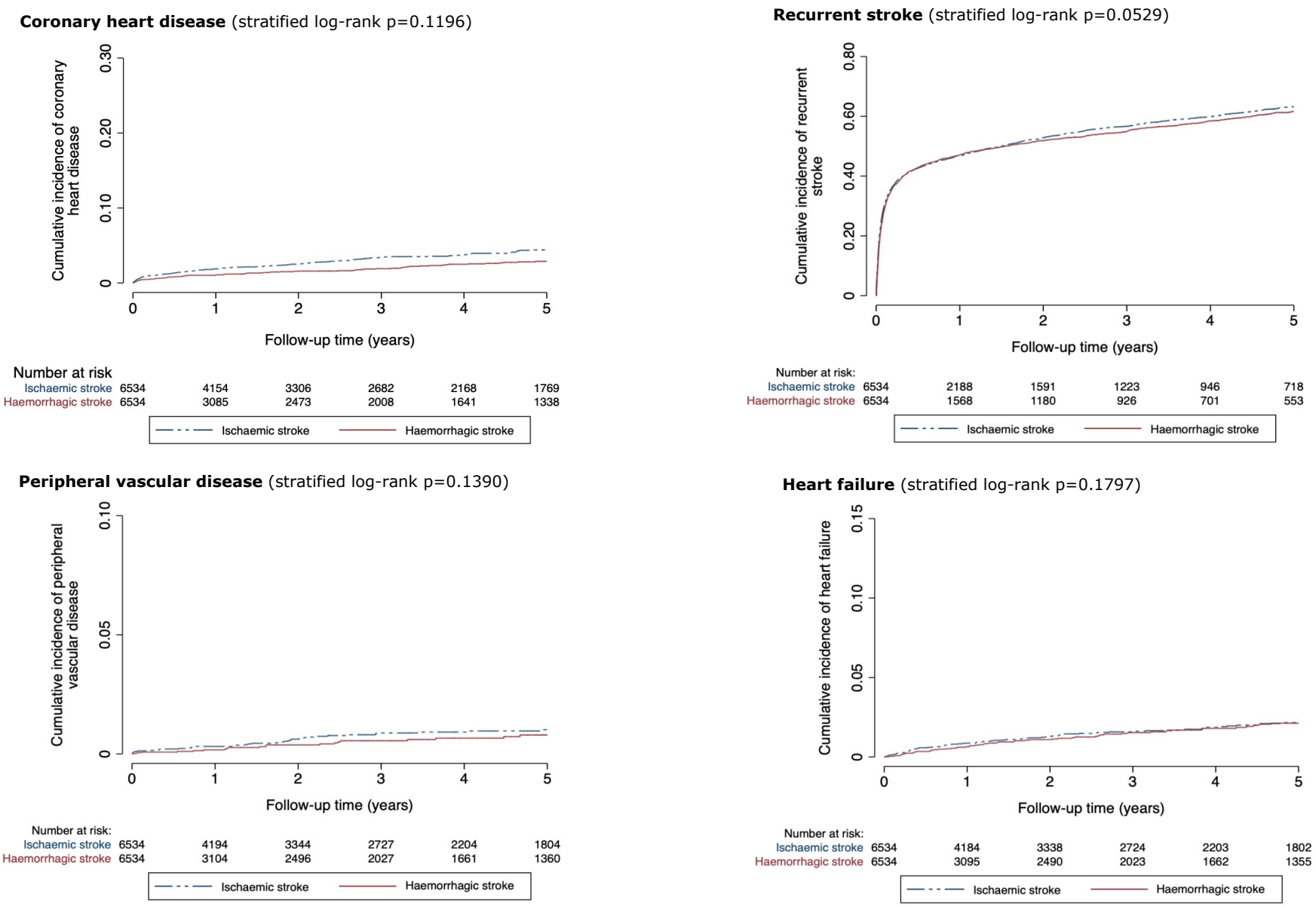
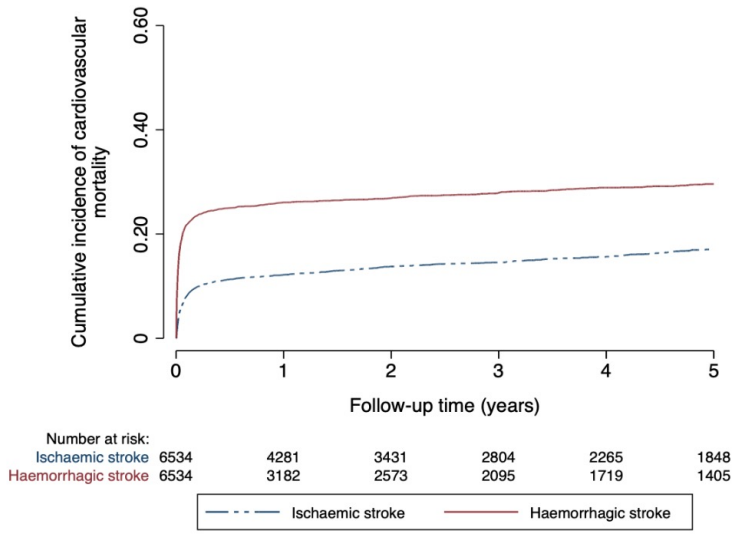
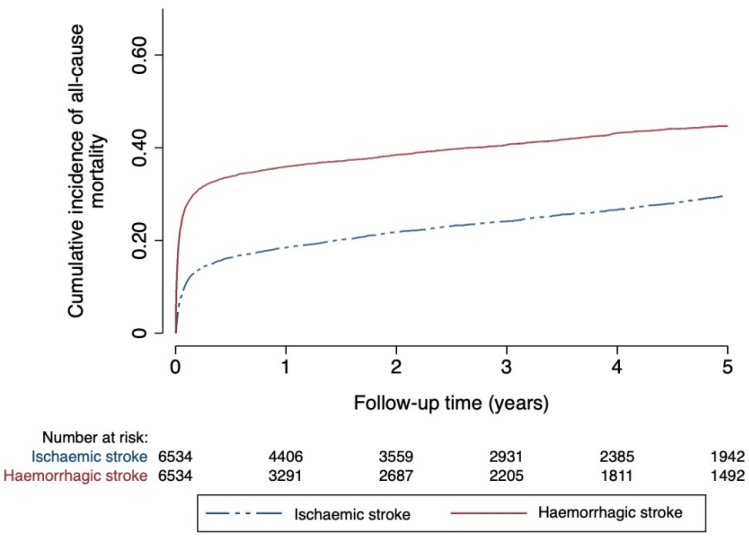


Figure 5.4 **Cumulative incidence plot for subsequent cardiovascular morbidity and mortality outcomes for the propensity-score matched cohort** (n=13,068)

Cardiovascular-related mortality (log rank p<0.0001)



All-cause mortality (log rank p<0.0001)



Composite MACE (log rank p<0.0001)

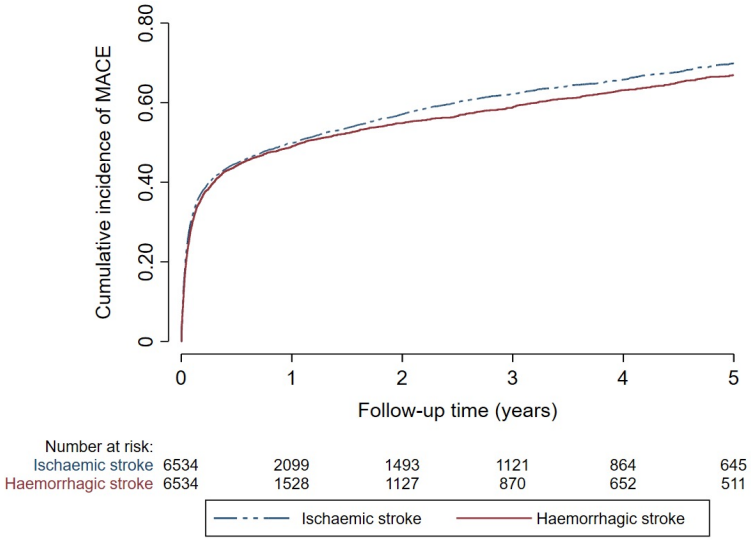


Figure 5.4 **Cumulative incidence plot for subsequent cardiovascular morbidity and mortality outcomes for the propensity-score matched cohort** (n=13,068)

Landmark-analysis

In the landmark analyses at 3 and 6 months, 17,193 patients with subsequent outcomes occurring within 3 months and 19,021 within 6 months of incident stroke events were excluded, respectively. Although the risk of subsequent mortality outcomes remained higher in patients with haemorrhagic stroke compared with ischaemic stroke patients, it was attenuated for both 3- and 6-month analyses – Table 5.5.

Table 5.5 Landmark analysis at 3 and 6 months for subsequent cardiovascular mortality according to incident stroke sub-type for the entire cohort with imputed values

Outcomes	3 months landmark analysis (n=14,898)				6 months landmark analysis (n=13,070)			
	Entire cohort 14,898 (100%)	**Ischaemic** 12,289 (82.5%)	**Haemorrhagic** 2,609 (17.5%)	***p*-value**	**Cohort** 13,070 (100%)	**Ischaemic** 1,039	**Haemorrhagic** 1,039	***p*-value**
Cardiovascular-related mortality								
Number (percent)	1,651 (11.1)	1,386 (11.3)	265 (10.2)	0.098	1,364 (10.4)	1,148 (10.6)	216 (9.5)	0.094
Follow-up time	2.12 (0.74 – 4.57)	2.17 (0.75 – 4.55)	2.02 (0.70 – 4.68)	0.4117	2.91 (1.37 – 5.11)	2.93 (1.42 – 5.08)	2.83 (1.10 – 5.52)	0.5865
Hazard ratio (95% CI)	–	Reference	1.19 (1.04 – 1.35)	0.011		Reference	1.17 (1.01 – 1.35)	0.036
All-cause mortality								
Number (percent)	4,723 (31.7)	3,919 (31.9)	804 (30.8)	0.284	4,079 (31.2)	3,399 (31.5)	680 (29.8)	0.106
Follow-up time	2.52 (0.93 – 5.15)	2.55 (0.96 – 5.11)	2.41 (0.80 – 5.27)	0.2997	3.08 (1.49 – 5.65)	3.07 (1.50 – 5.64)	3.17 (1.43 – 5.78)	0.0001
Hazard ratio (95% CI)	–	Reference	1.19 (1.10 – 1.29)	<0.001		Reference	1.16 (1.07 – 1.26)	<0.001

Follow-up time: Time from incident stroke event to mortality reported as median with interquartile range.

CI, confidence interval; HR, hazard ratio

Model adjusted for age at the time of incident stroke, sex, socioeconomic status, smoking status, body mass index, blood pressure, cholesterols (high-density lipoprotein, low-density lipoprotein, and total), diagnosis of an alcohol problem, atrial fibrillation, cancer, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, a prescription of antihypertensive, anticoagulant, antidiabetic, and potency of prescribed statin.

5.5 Discussion

Within a large population-based cohort with a long follow-up period, this study indicates that the risk of subsequent cardiovascular morbidity (CHD, recurrent stroke, PVD, and heart failure) was similar between patients with incident haemorrhagic or ischaemic stroke. Also, found a significantly increased risk of subsequent mortality outcomes (CVD-related and all-cause) in patients with incident haemorrhagic stroke as compared to individuals with incident ischaemic stroke.

This study is the first large-scale population-based study to compare long-term cardiovascular prognosis between patients with ischemic or haemorrhagic stroke over a long follow-up and shows that the risk of subsequent cardiovascular events in patients with haemorrhagic stroke is similar to that in patients with ischemic stroke. However, in a study analysing data from 4 population-based cohort studies in the USA, the rate of arterial ischaemic events (i.e., ischaemic stroke and myocardial infarction) was found to be 2–3 times higher in individuals with previous haemorrhagic stroke compared to those without.[233]

The finding of higher cardiovascular and all-cause mortality in patients with haemorrhagic stroke compared to ischemic stroke is consistent with previous studies.[236,237] A plausible explanation of this finding is that haemorrhagic strokes are usually more severe than ischemic strokes, given that stroke severity is a major predictor of stroke mortality.[238] To minimise the potential impact of incident stroke severity on subsequent mortality estimates in this study, I performed two landmark analyses at 3 and 6 months; the attenuation of mortality risk between the 3- and 6-month landmark analyses seems to support the assertion that stroke severity impacts on subsequent stroke-related mortality.

The strength of this study is in the size and representativeness of the CPRD GOLD dataset[87], this large population-based study used linked primary care, hospital, and mortality records to compare differences in subsequent cardiovascular

outcomes in the stroke subtypes. Additionally, the use of an incident cohort reflects current practice and avoids the distorting influences of bias present in cohorts with prevalent major adverse cardiovascular events. There are limitations in this study that should be taken into consideration. Although multiple confounders were accounted for in the multivariable analyses performed for the entire cohort and also in the propensity score matching, there may have been other residual confounders that could have influenced the overall results of this study. The severity of incident stroke was not available in the electronic health records and hence, it was not accounted for in the analyses. However, the landmark analyses at both 3 and 6 months after the incident stroke event were done to mitigate the effect of stroke severity. The proportional hazard assumption is the main premise for the Cox proportional hazard regression/model. Specifically, the model assumes that the hazard of each covariate does not change over time. Due to the large study population, the proportional hazard assumption is likely to be violated (i.e., a significant Schoenfeld residuals test). If the violation of proportionality is not too extreme, a single hazard ratio can still be a reasonable summary of the data.[239] The proportional hazards assumption was, therefore, not assessed in this study. The reporting of hazard ratios was, however, supplemented with the reporting of incident rates for outcomes within the groups.

Non-pharmaceutical interventions like weight reduction, reduction of salt intake, smoking cessation, and implementation of healthy dietary patterns constitute the cornerstone of a holistic preventive approach. Additionally, optimisation of pharmaceutical management of cardiovascular risk factors like hypertension, high cholesterol levels, diabetes mellitus, and strategies to increase patient adherence and persistence to it, is of paramount importance to reduce overall cardiovascular risk. Anti-thrombotic treatment is a challenging issue in patients with haemorrhagic stroke as the associated bleeding risk might counterbalance some of the conferred benefits[240]; ongoing studies assess the efficacy and safety of

different antithrombotic strategies in patients with previous intracranial haemorrhage.[241] In addition, new classes of anti-thrombotic are being developed like the FXIa inhibitors which showed to have a promising safety profile in preliminary reports.[242] Moreover, lipid-lowering treatment (including statins, ezetimibe and PCSK9 inhibitors) is crucial for cardiovascular risk reduction,[243–246] however, statins seem to increase the risk of haemorrhagic stroke in a dose-dependent manner, whereas PCSK9 inhibitors do not.[247] This implies that perhaps PCSK9 inhibitors may be a preferred lipid-lowering class in patients with previous haemorrhagic stroke.[247]

5.6 Conclusion

The results of this large population-based study of patients with incident haemorrhagic or ischaemic stroke suggest that patients with previous haemorrhagic stroke should be regarded as a very high-risk population for future cardiovascular events, as their risk is similar to patients with previous ischaemic stroke. Given that approximately 2.9 million individuals worldwide have a haemorrhagic stroke annually,[248] there is an urgent need for optimization of currently available strategies and development of new ones aiming to reduce the overall cardiovascular risk in this very-high risk population.

Summary

This chapter compared the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic stroke and those with incident ischaemic stroke. The next chapter examines the relationship between body mass index and subsequent MACE outcomes in patients with any type of incident stroke.

Chapter 6

Obesity and long-term outcomes after incident stroke: a prospective population-based cohort study

The previous chapter compared the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke. Obesity, a risk factor for stroke and other stroke-related risk factors (hypertension and diabetes), is commonly measured using body mass index. This chapter, therefore, assesses the relationship between BMI and subsequent MACE outcomes in patients with incident stroke.

A paper based on this research study has been published in the *Journal of Cachexia, Sarcopenia and Muscle*:

Akyea, R. K., Döhner, W., Iyen, B., Weng, S. F., Qureshi, N., & Ntaios, G. (2021). Obesity and long-term outcomes after incident stroke: a prospective population-based cohort study. Journal of Cachexia, Sarcopenia and Muscle, https://doi.org/10.1002/jcsm.12818

6.1 Abstract

Background: The association between obesity, major adverse cardiovascular events (MACE), and mortality in patients with incident stroke is not well established. This study assessed the relationship between body mass index (BMI) and MACE in patients with incident stroke.

Methods: This cohort study identified 30,702 patients aged ≥18 years from UK CPRD GOLD and HES data with incident stroke between Jan-1998 and Dec-2017, a BMI recorded within 24 months before the incident stroke and no prior history of MACE. BMI was categorised as underweight (<18.5 kg/m^2), normal (18.5 – 24.9 kg/m^2), overweight (25.0 – 29.9 kg/m^2), obesity class I (30.0 – 34.9 kg/m^2), class II (35.0 – 39.9 kg/m^2) and class III (≥40 kg/m2). Multivariable Cox regression was used to assess differences in MACE risk between BMI categories.

Results: At baseline, 1,217 (4.0%) were underweight, 10,783 (35.1%) had a normal BMI, 10,979 (35.8%) had overweight, 5,206 (17.0%) had obesity class I, 1,749 (5.7%) class II, and 768 (2.5%) class III. In multivariable analysis, higher BMI categories were associated with lower risk of subsequent outcomes:

- MACE [overweight (BMI: 25.0-29.9 kg/m^2): HR 0.96, 95% CI 0.93 – 0.99)],
- PVD [overweight: HR 0.65, 95% CI 0.49 – 0.85; obesity class III (BMI: ≥40 kg/m^2): HR 0.19, 95% CI 0.50 – 0.77],
- CVD-related mortality [overweight: HR 0.80, 95% CI 0.74 – 0.86; obesity class I (BMI: 30.0-34.9 kg/m^2): HR 0.79, 95% 0.71 – 0.88; class II (BMI: 35.0-39.9 kg/m^2): HR 0.80, 95% CI 0.67 – 0.96]; and
- all-cause mortality [overweight: HR 0.75, 95% CI 0.71 – 0.79; obesity class I: HR 0.75, 95% CI 0.70 – 0.81; class II: HR 0.77, 95% CI 0.68 – 0.86]

when compared to those with normal BMI. The results were similar irrespective of sex, diabetes mellitus, smoking or cancer at the time of incident stroke.

Conclusions: In patients with incident stroke, overweight or obesity were associated with a more favourable prognosis for subsequent MACE, PVD, and mortality, irrespective of sex, diabetes mellitus, smoking or cancer at baseline.

6.2 Introduction

Obesity is an established risk factor for stroke,[249] but the association of increased body mass index (BMI) with survival after stroke remains contentious. Contrary to evidence in the general population,[250] in patients with established cardiovascular disease (CVD) increased BMI is independently associated with better outcome.[251–254] Many studies have shown that increased BMI has a protective effect on survival after stroke,[255,256] while other studies have not confirmed an obesity paradox in patients with stroke.[257] The association between BMI and composite major adverse cardiovascular event (MACE) and its constituent outcomes have, however, not been studied using a population-based cohort in patients with any subtype of incident stroke.

Using a large population-based cohort in the United Kingdom, this study aimed to examine the relationship between BMI and MACE outcomes during long-term follow-up in patients with any subtype of incident stroke.

6.3 Methods

Data source

This prospective population-based cohort study used the UK CPRD GOLD[87] linked to HES APC[208], ONS death registry[95], and social deprivation data.[97] The databases have been previously described in Chapter 2 (Section 2.2).

Study population

The study cohort of patients with the first record of non-fatal stroke in either CPRD GOLD or HES APC between 1 January 1998 and 31 December 2017 has been previously described in Chapter 2 (Section 2.3.1).

Cohort demographics and baseline characteristics

Age was defined at the time of the incident stroke. Ethnicity was categorised into six groups: Asian, Black, Mixed, Other, White and unknown.[90] Socioeconomic status based on the English IMD 2015,[97] described in Chapter 2 (Section 2.2.4), was categorised into quintiles (quintile 1 – least deprived group to quintile 5 – most deprived group). Medication prescriptions (issue of a prescription) at baseline was defined as a prescription within 12 months before the incident stroke. For cholesterol (low-density lipoprotein (LDL), high-density lipoprotein (HDL) and total), BMI, blood pressure measures (diastolic and systolic), and glomerular filtration rate (GFR), the most recent values/measures within 24 months before incident stroke were used. All other comorbidities were defined based on the latest record before the incident stroke.

Body mass index

BMI was categorised according to the WHO criteria as underweight (BMI <18.5 kg/m^2), normal weight (BMI 18.5-24.9 kg/m^2), overweight (BMI 25.0-29.9 kg/m^2), obesity class I (BMI 30.0-34.9 kg/m^2), obesity class II (BMI 35.0-39.9 kg/m^2), obesity class III (BMI $\geq$ 40 kg/m^2).[258] Accordingly, and in line with accumulating epidemiologic evidence, patients within the normal BMI category (18.5-24.9 kg/m^2) were used as the reference group.[259]

Outcome measures

The first subsequent MACE after incident stroke was the primary outcome. MACE was defined as a composite of new-onset CHD, recurrent stroke, PVD, heart

failure, or cardiovascular-related mortality, based on the record from across the linked data sources (CPRD, HES or ONS registry). All-cause mortality was considered as a secondary outcome.

Statistical analysis

The Shapiro-Wilk test was used to assess the normality of distribution for continuous variables.[260] Kruskal-Wallis test for continuous data and the chi-squared test for categorical data were used to compare baseline characteristics between BMI categories. The level of missing values ranged between 3.1% for blood pressure measures to 57.4% for GFR. Details on the proportion of missingness are provided in Appendix F.6.1. To estimate missing values for BMI, systolic and diastolic blood pressures, GFR, HDL-C, LDL-C and total cholesterol levels, multiple imputation by chained equations was used to generate imputed datasets as described in Chapter 2 (Section 2.3.4).[108–110] The imputed datasets were pooled into a single dataset using Rubin's rules.[111] Differences in baseline characteristics between those with and without a BMI record within 24 months of incident stroke is provided in the Appendix F.6.2. Event rates between BMI categories were analysed by multivariable Cox regression models using the category of normal BMI as reference. Time to event curves for BMI categories was made for MACE outcomes. Hazard ratios (HR) and 95% confidence intervals (95% CI) for the outcomes according to BMI category were calculated in Cox regression models adjusted for: (a) age and sex (b) age, sex, socioeconomic status, current smoking, history of alcohol problem, atrial fibrillation, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, prescription of angiotensin-converting enzyme (ACE) inhibitor, anti-hypertensive, anti-diabetic, anti-platelet, beta-blocker, calcium channel blocker, non-steroidal anti-inflammatory drugs (NSAIDs), statin potency, diastolic and systolic blood pressure, GFR, total cholesterol (full adjustment model). A restricted cubic spline with 3-5 knots [lowest Akaike information criterion (AIC)] was used for the non-

linear relationship between BMI and outcomes. Subgroup analyses according to sex, a diagnosis of diabetes, current smoking status, and a diagnosis of cancer (excluding those with a cancer diagnosis) at the time of incident stroke was done to explore any potential reverse causality pathways. All statistical analyses were performed using Stata SE version 16.1 (StataCorp LP). An alpha level of 0.05 was used for all analyses to define statistical significance.

6.4 Results

A total of 30,702 individuals with baseline BMI records (53% women) were included in this study. The median age for the study cohort was 75 years (IQR: 65–82). The distribution of BMI within the study cohort is present in Figure 6.1. Most of the individuals were within the overweight and obesity categories (60.9%) and 35.1% had normal BMI. Clinical characteristics and medications prescribed at baseline across the BMI categories are presented in Table 6.1 and by sex in Appendix F.6.3. Individuals in the obese classes (I-III) were younger and had a higher prevalence of hypertension and diabetes mellitus at baseline.

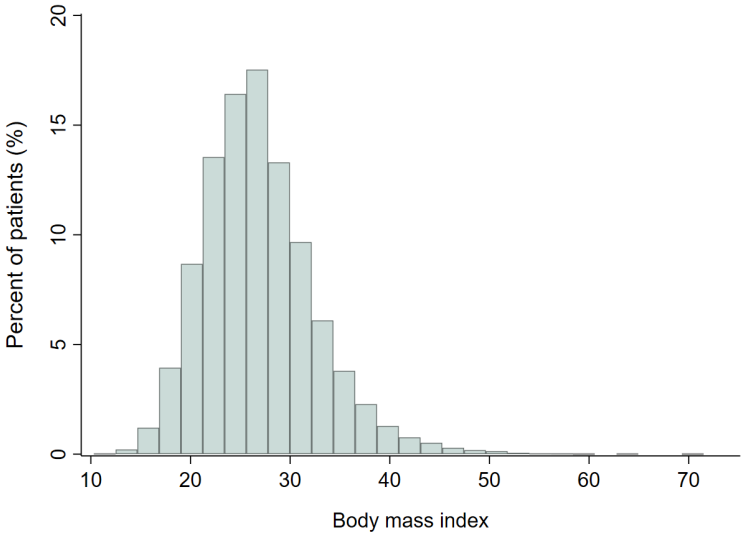


Figure 6.1 Distribution of body mass index in the study population

Table 6.1 Characteristics of the study population at the time of incident stroke according to body mass index categories

Characteristics	**< 18.5** 1,217 (4.0%)	**18.5 – 24.9** 10,783 (35.1%)	**25.0 – 29.9** 10,979 (35.8%)	**30.0 – 34.9** 5,206 (17.0%)	**35.0 – 39.9** 1,749 (5.7%)	**≥ 40 kg/m²** 768 (2.5%)	***p*-value**
Follow-up, median (IQR)	10.2 (5.7–15.4)	12.1 (7.2–16.6)	13.4 (8.5–17.5)	13.8 (8.8–17.8)	13.3 (8.7–17.5)	13.8 (8.2–17.7)	0.0001
Females	935 (76.8)	6,144 (57.0)	5,164 (47.0)	2,620 (50.3)	1,025 (58.6)	511 (66.5)	<0.001
Age (years), median (IQR)	81 (72–87)	78 (69–85)	74 (65–81)	71 (62–79)	68 (58–75)	63.5 (53–72)	0.0001
Incident stroke subtype							<0.001
Haemorrhagic stroke	146 (12.0)	1,037 (9.6)	869 (7.9)	383 (7.4)	139 (8.0)	54 (7.0)	
Ischaemic stroke	437 (35.9)	3,996 (37.1)	4,204 (38.3)	2,071 (39.8)	703 (40.2)	323 (42.1)	
Stroke (not specified)	634 (52.1)	5,750 (53.3)	5,906 (53.8)	2,752 (52.9)	907 (51.9)	391 (50.9)	
Ethnicity							<0.001
Asian	18 (1.5)	191 (1.8)	197 (1.8)	84 (1.6)	23 (1.3)	4 (0.7)	
Black	9 (0.7)	89 (0.8)	105 (1.0)	86 (1.7)	39 (2.2)	20 (2.6)	
Mixed	0	12 (0.1)	13 (0.1)	11 (0.2)	3 (0.2)	3 (0.4)	
Other	6 (0.5)	75 (0.7)	80 (0.7)	38 (0.7)	13 (0.7)	4 (0.5)	
White	1,116 (91.7)	9,758 (90.5)	10,031 (91.4)	4,748 (91.2)	1,598 (91.4)	704 (91.7)	
Unknown	68 (5.6)	658 (6.1)	553 (5.0)	239 (4.6)	73 (4.2)	32 (4.2)	
Socioeconomic status							<0.001
1 (Least deprived)	233 (19.2)	2,369 (22.0)	2,386 (21.7)	926 (17.8)	249 (14.2)	94 (12.2)	
2	257 (21.1)	2,368 (22.0)	2,443 (22.3)	1,124 (21.6)	368 (21.0)	143 (18.6)	
3	243 (20.0)	2,288 (21.2)	2,275 (20.7)	1,077 (20.7)	372 (21.3)	177 (23.1)	
4	272 (22.4)	1,987 (18.4)	2,124 (19.4)	1,062 (20.4)	374 (21.4)	165 (21.5)	
5 (Most deprived)	210 (17.3)	1,763 (16.4)	1,737 (15.8)	1,012 (19.4)	384 (22.0)	189 (24.6)	
Unknown	2 (0.2)	8 (0.1)	14 (0.1)	5 (0.1)	2 (0.1)	0	

Characteristics	**< 18.5** 1,217 (4.0%)	**18.5 – 24.9** 10,783 (35.1%)	**25.0 – 29.9** 10,979 (35.8%)	**30.0 – 34.9** 5,206 (17.0%)	**35.0 – 39.9** 1,749 (5.7%)	**≥ 40 kg/m²** 768 (2.5%)	***p*-value**
Current smokers	349 (28.7)	2,177 (20.2)	1,841 (16.8)	874 (16.8)	306 (17.5)	142 (18.5)	<0.001
DBP (mmHg)	78 (70–82)	79 (70–83)	80 (71–85)	80 (73–86)	80 (74–86)	80 (75–88)	<0.001
SBP (mmHg)	138 (122–149)	139 (128–148)	140 (130–150)	140 (130–150)	140 (130–150)	140 (130–150)	<0.001
HDL cholesterol (mmol/L)	1.8 (1.5–2.0)	1.6 (1.3–1.8)	1.4 (1.2–1.6)	1.3 (1.1–1.5)	1.2 (1.1–1.4)	1.1 (1.0–1.3)	0.0001
LDL cholesterol (mmol/L)	2.8 (2.5–3.2)	2.9 (2.5–3.3)	2.9 (2.4–3.4)	2.9 (2.3–3.4)	2.9 (2.3–3.4)	2.9 (2.4–3.5)	0.0001
Total cholesterol (mmol/L)	5.2 (4.6–5.6)	5.1 (4.5–5.5)	5.0 (4.4–5.6)	5.0 (4.3–5.6)	5.0 (4.3–5.6)	5.0 (4.4–5.5)	0.0001
GFR	67.5 (60.3–74.7)	67.0 (60.3–73.0)	67.4 (60.9–73.1)	68.0 (61.0–73.9)	68.7 (61.6–75.3)	69.2 (63.0–76.0)	0.0001
Comorbidities at baseline							
Alcohol problem	61 (5.0)	366 (3.4)	300 (2.7)	164 (3.2)	48 (2.7)	28 (3.7)	<0.001
Atrial fibrillation	154 (12.7)	1,273 (11.8)	1,159 (10.6)	483 (9.3)	166 (9.5)	74 (9.6)	<0.001
Cancer	256 (21.0)	2,273 (21.1)	1,956 (17.8)	818 (15.7)	227 (13.0)	97 (12.6)	<0.001
Chronic kidney disease	163 (13.4)	1,491 (13.8)	1,601 (14.6)	824 (15.8)	281 (16.1)	102 (13.3)	0.005
Diabetes mellitus	121 (9.9)	1,672 (15.5)	2,515 (22.9)	1,589 (30.5)	660 (37.7)	295 (38.4)	<0.001
Type-1 diabetes	11 (0.9)	142 (1.3)	174 (1.6)	105 (2.0)	40 (1.3)	23 (3.0)	<0.001
Type-2 diabetes	95 (7.8)	1,331 (12.3)	2,112 (19.2)	1,408 (27.1)	594 (34.0)	265 (34.5)	<0.001
Dyslipidaemia	76 (6.2)	1,215 (11.3)	1,569 (14.3)	846 (16.3)	291 (16.6)	124 (16.2)	<0.001
Hypertension	528 (43.4)	5,555 (51.5)	6,402 (58.3)	3,359 (64.5)	1,164 (66.6)	503 (65.5)	<0.001
Transient ischaemic attack	263 (21.6)	2,472 (22.9)	2,629 (24.0)	1,167 (22.4)	336 (19.2)	131 (17.1)	<0.001
Prescribed medications at baseline							
ACE inhibitor	287 (23.6)	3,518 (32.6)	4,455 (40.6)	2,637 (50.7)	958 (54.8)	429 (55.9)	<0.001
Anti-diabetic	73 (6.0)	1,272 (11.8)	2,025 (18.4)	1,326 (25.5)	556 (31.8)	254 (33.1)	<0.001
Anti-hypertensive	550 (45.2)	5,732 (53.2)	6,759 (61.6)	3,587 (68.9)	1,260 (72.0)	545 (71.0)	<0.001
Antiplatelets	495 (40.7)	4,605 (42.7)	4,806 (43.8)	2,286 (43.9)	745 (42.6)	310 (40.4)	0.090

Characteristics	**< 18.5** 1,217 (4.0%)	**18.5 – 24.9** 10,783 (35.1%)	**25.0 – 29.9** 10,979 (35.8%)	**30.0 – 34.9** 5,206 (17.0%)	**35.0 – 39.9** 1,749 (5.7%)	**≥ 40 kg/m²** 768 (2.5%)	***p*-value**
Beta-blockers	265 (21.8)	2,604 (24.2)	2,996 (27.3)	1,480 (28.4)	560 (32.0)	246 (32.0)	<0.001
Calcium channel blocker	280 (23.0)	2,873 (26.6)	3,469 (31.6)	1,789 (34.4)	657 (37.6)	264 (34.4)	<0.001
NSAIDS	234 (19.2)	2,508 (23.3)	3,030 (27.6)	1,675 (32.2)	598 (34.2)	284 (37.0)	<0.001
Statin							<0.001
Low intensity	39 (3.2)	477 (4.4)	596 (5.4)	281 (5.4)	107 (6.1)	36 (4.7)	
Moderate intensity	210 (17.3)	2,370 (22.0)	3,115 (28.4)	1,627 (31.3)	567 (32.4)	240 (31.3)	
High intensity	32 (2.6)	524 (4.9)	755 (6.9)	488 (9.4)	188 (10.8)	87 (11.3)	

Nutritional status for the body mass index categories (kg/m^2): underweight (< 18.5); normal weight (18.5–24.9); pre-obese (25.0–29.9); obesity class I (30.0–34.9); obesity class II (35.0–39.9); obesity class III (≥40).

ACE: angiotensin-converting enzyme; DBP: diastolic blood pressure; GFR: glomerular filtration rate; HDL: high-density lipoprotein; IQR: interquartile range; LDL: low-density lipoprotein; n: frequency/numbers; NSAIDs: non-steroidal anti-inflammatory drug; SBP: systolic blood pressure; %: per cent.

During a median follow-up of 12.9 years (IQR: 7.9–17.2 years), 20,881 (68.0%) individuals had a subsequent MACE outcome recorded. The proportion of subsequent MACE outcomes was similar across the BMI categories. Table 6.2 details the number and proportion for all the MACE and all the individual constituent outcomes.

In multivariable analysis, individuals within higher BMI categories were associated with lower risk of subsequent outcomes:

- MACE [overweight (BMI: 25.0-29.9 kg/m^2): HR 0.96, 95% CI 0.93 – 0.99)],
- PVD [overweight: HR 0.65, 95% CI 0.49 – 0.85; obesity class III (BMI: ≥40 kg/m^2): HR 0.19, 95% CI 0.50 – 0.77],
- CVD-related mortality [overweight: HR 0.80, 95% CI 0.74 – 0.86; obesity class I (BMI: 30.0-34.9 kg/m^2): HR 0.79, 95% 0.71 – 0.88; class II (BMI: 35.0-39.9 kg/m^2): HR 0.80, 95% CI 0.67 – 0.96]; and
- all-cause mortality [overweight: HR 0.75, 95% CI 0.71 – 0.79; obesity class I: HR 0.75, 95% CI 0.70 – 0.81; class II: HR 0.77, 95% CI 0.68 – 0.86]

when compared to those within the normal BMI category – Table 6.3.

Appendices F.6.4 – F.6.6 present the results disaggregated by sex, diabetes mellitus, and smoking status at the time of incident stroke, respectively. Table 6.4 and Appendix F.6.7 presents similar results after excluding 5,627 (18.3%) individuals with a cancer diagnosis at baseline and excluding 8,735 (28.5%) individuals with first subsequent outcomes within 30 days of incident stroke, respectively. The Kaplan-Meier curves for MACE and all-cause mortality across the BMI categories over a 10-year follow-up period is presented in Figure 6.2.

Table 6.2 Number and proportion of first subsequent outcomes within the body mass index categories

Outcomes	< 18.5 n=1,217 (4.0%)	18.5 – 24.9 n=10,783 (35.1%)	25.0 – 29.9 n=10,979 (35.8%)	30.0 – 34.9 n=5,206 (17.0%)	35.0 – 39.9 n=1,749 (5.7%)	≥ 40 kg/m² n=768 (2.5%)	*p*-value
Composite MACE							
Follow-up time	0.14 (0.03-1.10)	0.21 (0.03-1.46)	0.27 (0.04-1.77)	0.23 (0.03-1.66)	0.19 (0.03-1.67)	0.16 (0.03-1.28)	0.0001
Number of events (percent)	806 (66.2)	7,326 (67.9)	7,497 (68.3)	3,545 (68.1)	1,217 (69.6)	490 (63.8)	0.064
CHD							
Follow-up time	0.83 (0.28-2.20)	1.49 (0.31-3.42)	1.91 (0.60-4.59)	1.76 (0.52-3.86)	2.89 (1.03-4.62)	1.62 (0.90-3.79)	0.0001
Number of events (percent)	24 (2.0)	378 (3.5)	459 (4.2)	252 (4.8)	86 (4.9)	27 (3.5)	<0.001
Recurrent stroke							
Follow-up time	0.15 (0.04-1.03)	0.18 (0.03-1.24)	0.19 (0.03-1.29)	0.15 (0.02-1.15)	0.11 (0.03-1.02)	0.10 (0.02-1.05)	0.0001
Number of events (percent)	490 (40.3)	5,119 (47.5)	5,580 (50.8)	2,636 (50.6)	908 (51.9)	379 (49.4)	<0.001
PVD							
Follow-up time	1.83 (1.07-2.76)	1.22 (0.50-2.89)	1.58 (0.77-4.63)	2.13 (0.73-4.67)	2.26 (1.22-3.74)	4.93 (1.23-8.63)	0.2636
Number of events (percent)	17 (1.4)	114 (1.1)	97 (0.9)	61 (1.2)	19 (1.1)	2 (0.3)	0.087
Heart failure							
Follow-up time	1.56 (0.64-3.35)	1.23 (0.32-3.47)	2.33 (0.74-4.64)	2.12 (0.54-5.62)	2.14 (0.59-4.70)	2.68 (1.30-5.84)	0.004
Number of events (percent)	20 (1.6)	209 (1.9)	241 (2.2)	136 (2.6)	62 (3.5)	20 (2.6)	<0.001
Cardiovascular mortality							
Follow-up time	0.07 (0.02-0.83)	0.09 (0.02-1.37)	0.10 (0.02-1.95)	0.12 (0.02-2.15)	0.16 (0.02-2.64)	0.09 (0.01-0.90)	0.0797
Number of events (percent)	255 (21.0)	1,506 (14.0)	1,120 (10.2)	460 (8.8)	142 (8.1)	62 (8.1)	<0.001
All-cause mortality							
Follow-up time	0.35 (0.4-2.29)	0.68 (0.06-3.25)	0.84 (0.06-4.17)	1.05 (0.06-4.14)	0.69 (0.06-4.18)	0.43 (0.04-2.38)	0.0001
Number of events (percent)	573 (47.1)	3,421 (31.7)	2,557 (23.3)	1,053 (20.2)	314 (18.0)	140 (18.2)	<0.001

CHD: coronary heart disease; MACE: major adverse cardiovascular event; PVD: peripheral vascular disease.

Follow-up time: Time from incident stroke event to first subsequent event reported as median with interquartile range in years.

Table 6.3 Outcomes in body mass index subgroups

Outcomes	**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m²** n=768 (2.5%)
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Composite MACE					
Age and sex adjusted	1.14 (1.06 – 1.23)	0.96 (0.93 – 0.99)	0.98 (0.95 – 1.03)	1.10 (1.04 – 1.17)	1.07 (0.97 – 1.17)
Full adjustment	1.12 (1.05 – 1.21)	0.96 (0.93 – 0.99)	0.98 (0.94 – 1.02)	1.08 (1.01 – 1.15)	1.04 (0.95 – 1.14)
CHD					
Age and sex adjusted	0.81 (0.53 – 1.22)	1.00 (0.87 – 1.15)	1.20 (1.02 – 1.41)	1.29 (1.01 – 1.63)	1.01 (0.68 – 1.51)
Full adjustment	0.85 (0.56 – 1.29)	0.94 (0.82 – 1.09)	1.05 (0.89 – 1.24)	1.06 (0.83 – 1.35)	0.82 (0.55 – 1.23)
Recurrent stroke					
Age and sex adjusted	1.01 (0.92 – 1.10)	1.00 (0.97 – 1.04)	1.00 (0.96 – 1.05)	1.07 (1.00 – 1.15)	1.05 (0.94 – 1.16)
Full adjustment	1.00 (0.91 – 1.09)	1.01 (0.98 – 1.05)	1.02 (0.97 – 1.07)	1.09 (1.02 – 1.18)	1.06 (0.95 – 1.18)
PVD					
Age and sex adjusted	1.96 (1.17 – 3.26)	0.71 (0.54 – 0.93)	1.00 (0.73 – 1.37)	1.02 (0.62 – 1.66)	0.28 (0.07 – 1.14)
Full adjustment	1.91 (1.14 – 3.19)	0.65 (0.49 – 0.85)	0.79 (0.57 – 1.09)	0.70 (0.42 – 1.17)	0.19 (0.05 – 0.77)
Heart failure					
Age and sex adjusted	1.09 (0.69 – 1.74)	1.12 (0.93 – 1.35)	1.60 (1.28 – 1.99)	2.62 (1.96 – 3.50)	2.60 (1.63 – 4.15)
Full adjustment	1.13 (0.71 – 1.80)	1.05 (0.87 – 1.26)	1.41 (1.12 – 1.76)	2.10 (1.56 – 2.83)	1.97 (1.23 – 3.17)
Cardiovascular mortality					
Age and sex adjusted	1.57 (1.38 – 1.80)	0.80 (0.74 – 0.87)	0.82 (0.74 – 0.91)	0.89 (0.75 – 1.06)	1.16 (0.90 -1.50)
Full adjustment	1.53 (1.34 – 1.75)	0.80 (0.74 – 0.86)	0.79 (0.71 – 0.88)	0.80 (0.67 – 0.96)	1.02 (0.79 – 1.32)

Outcomes	**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m²** n=768 (2.5%)
All-cause mortality					
Age and sex adjusted	1.73 (1.58 – 1.89)	0.75 (0.71 – 0.79)	0.76 (0.70 – 0.81)	0.80 (0.71 – 0.90)	1.06 (0.89 – 1.26)
Full adjustment	1.64 (1.50 – 1.80)	0.75 (0.71 – 0.79)	0.75 (0.70 – 0.81)	0.77 (0.68 – 0.86)	0.99 (0.84 – 1.18)

CHD: coronary heart disease; HR: hazards ratio; MACE: major adverse cardiovascular event; PVD: peripheral vascular disease.

Full adjustment for age, sex, socioeconomic status, current smoking, history of an alcohol problem, atrial fibrillation, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, prescription of ACE inhibitor, anti-hypertensive, anti-diabetic, anti-platelet, beta-blocker, calcium channel blocker, NSAIDS, statin potency, diastolic and systolic blood pressure, glomerular filtration rate, total cholesterol.

Reference category: Normal weight patients with a BMI of 18.5-24.9 kg/m^2.

Table 6.4 Outcomes in body mass index subgroups excluding patients with cancer at the time of incident stroke (n=25,075)

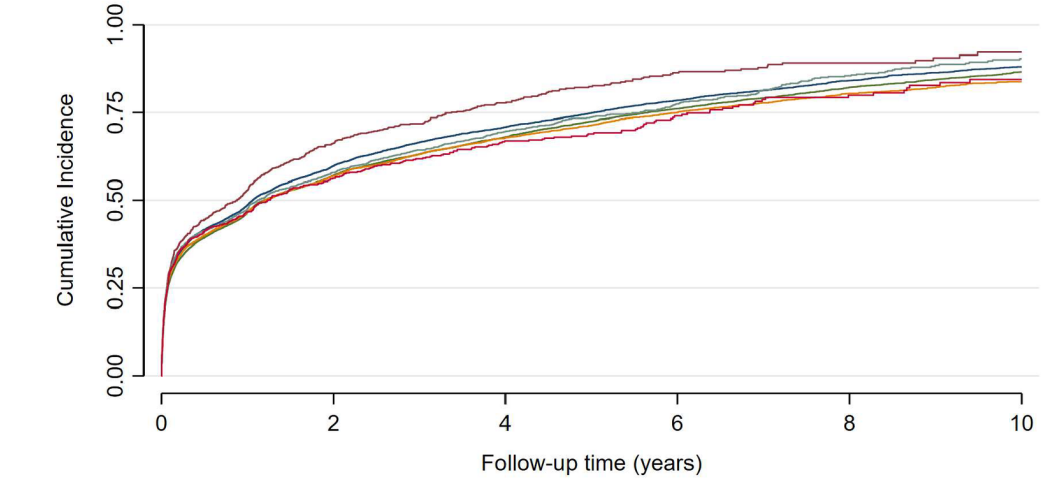
Outcomes	**< 18.5** n=961 (3.8%)	**25.0 – 29.9** n=9,023 (36.0%)	**30.0 – 34.9** n=4,388 (17.5%)	**35.0 – 39.9** n=1,522 (6.1%)	**≥ 40 kg/m²** n=671 (2.7%)
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Composite MACE					
Full adjustment	1.13 (1.04 – 1.22)	0.97 (0.93 – 1.00)	0.97 (0.93 – 1.02)	1.07 (1.00 – 1.15)	1.06 (0.96 – 1.17)
CHD					
Full adjustment	0.99 (0.65 – 1.52)	0.95 (0.82 – 1.11)	1.05 (0.87 – 1.25)	1.06 (0.82 – 1.37)	0.84 (0.55 – 1.27)
Recurrent stroke					
Full adjustment	1.00 (0.90 – 1.11)	1.02 (0.98 – 1.07)	1.02 (0.97 – 1.08)	1.10 (0.12 – 1.19)	1.08 (0.97 – 1.21)
PVD					
Full adjustment	1.92 (1.09 – 3.40)	0.69 (0.51 – 0.92)	0.73 (0.51 – 1.05)	0.72 (0.43 – 1.22)	0.21 (0.05 – 0.85)
Heart failure					
Full adjustment	0.89 (0.49 – 1.60)	1.05 (0.85 – 1.30)	1.46 (1.14 – 1.87)	2.12 (1.53 – 2.94)	2.11 (1.27 – 3.50)
Cardiovascular mortality					
Full adjustment	1.48 (1.27 – 1.72)	0.78 (0.71 – 0.85)	0.77 (0.68 – 0.86)	0.77 (0.64 – 0.94)	1.10 (0.84 – 1.45)
All-cause mortality					
Full adjustment	1.68 (1.52 – 1.86)	0.74 (0.70 – 0.79)	0.73 (0.67 – 0.80)	0.75 (0.66 – 0.86)	1.03 (0.85 – 1.25)

CHD: coronary heart disease; HR: hazards ratio; MACE: major adverse cardiovascular event; PVD: peripheral vascular disease.

Full adjustment for age, sex, socioeconomic status, current smoking, history of an alcohol problem, atrial fibrillation, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, prescription of ACE inhibitor, anti-hypertensive, anti-diabetic, anti-platelet, beta-blocker, calcium channel blocker, NSAIDS, statin potency, diastolic and systolic blood pressure, glomerular filtration rate, total cholesterol.

Reference category: Normal weight patients with a BMI of 18.5-24.9 kg/m^2.

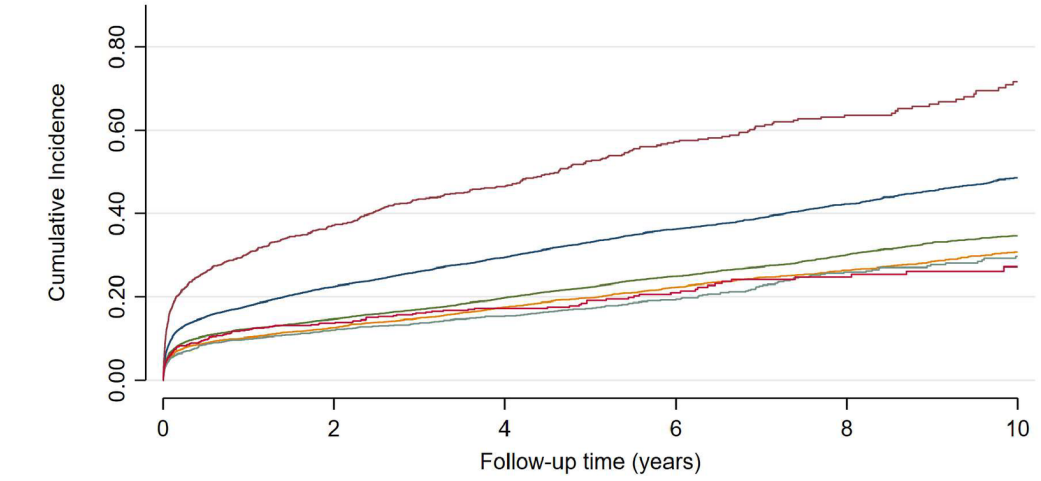
(a) Major adverse cardiovascular event



Number at risk:

BMI 18.5-24.9	10783	3025	1620	852	426	221
BMI < 18.5	1217	239	112	40	19	7
BMI 25.0-29.9	10979	3506	2036	1129	557	254
BMI 30.0-34.9	5206	1673	945	555	316	172
BMI 35.0-39.9	1749	536	299	143	64	29
BMI > 40.0	768	240	133	64	31	14

(b) All-cause mortality



Number at risk:

BMI 18.5-24.9	10783	6021	3921	2406	1420	746
BMI < 18.5	1217	524	311	147	80	37
BMI 25.0-29.9	10979	7083	4919	3232	1976	1064
BMI 30.0-34.9	5206	3360	2291	1483	914	512
BMI 35.0-39.9	1749	1099	768	476	288	158
BMI > 40.0	768	468	318	192	116	64

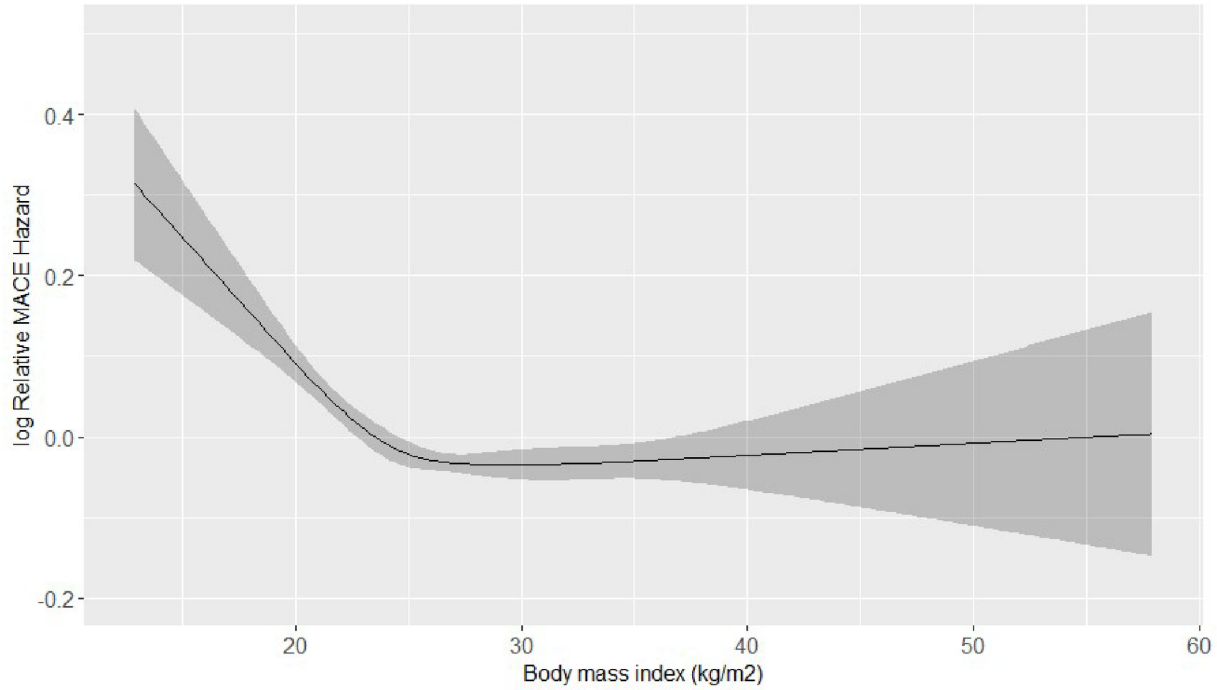
Figure 6.2 Kaplan-Meier plots for MACE and all-cause mortality outcomes

When compared with normal BMI, underweight was associated with a higher risk of MACE [HR 1.12 (95% CI 1.05-1.21)], PVD [HR 1.91 (95% CI 1.14-3.19)], cardiovascular-related death [HR 1.53 (95% CI 1.34-1.75)], and all-cause mortality [HR 1.64 (95% CI 1.50-1.80)].

Individuals who were obese had a higher risk of subsequent heart failure [obesity class I: HR 1.41 (95% CI 1.12-1.76); obesity class II: HR 2.10 (95% CI 1.56-2.83); obesity class III: HR 1.97 (95% CI 1.23-3.17)] when compared with those with a normal BMI.

The association between BMI and subsequent MACE outcome as well as all-cause mortality was non-linear as shown by the restricted cubic splines, Figure 6.3. The risk for both subsequent MACE and all-cause mortality outcomes were significantly higher at lower BMI and lower from BMI greater than 25kg/m^2.

(a) Major adverse cardiovascular event



(b) All-cause mortality

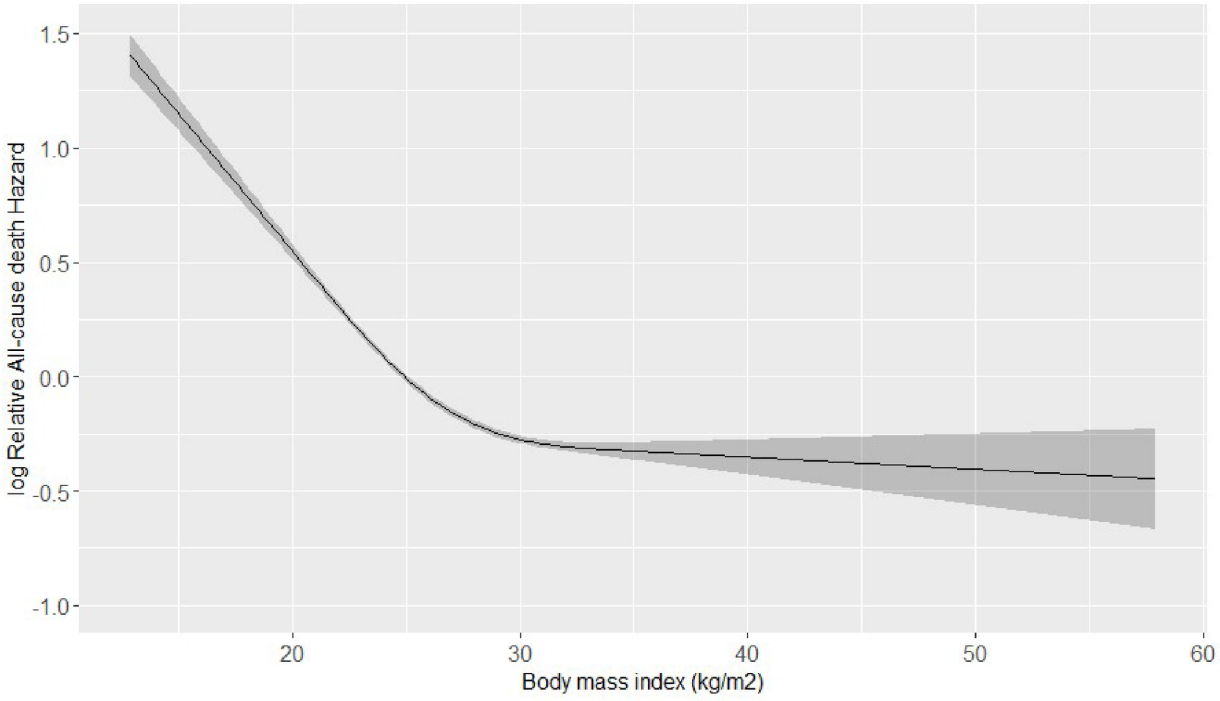


Figure 6.3 Restricted cubic splines for the association between body mass index (continuous variable) and outcomes

6.5 Discussion

In this prospective population-based cohort study of 30,702 patients with incident stroke followed for a median duration of 12.9 years, overweight (BMI: 25.0-29.9 kg/m^2) or obesity (BMI: ≥30 kg/m^2) was associated with a more favourable prognosis for subsequent MACE, PVD, cardiovascular mortality and all-cause mortality, irrespective of sex, diabetes mellitus, smoking or cancer at the time of incident stroke.

After the first reports of the stroke-obesity paradox,[261] several confirmatory reports were subsequently published.[262] The stroke-obesity paradox comes in contrast to the well-established association between obesity and the risk of cardiovascular disease in the general population.[263] Different explanations were proposed to explain this paradoxical conclusion. It was suggested that this may simply represent an erroneous finding associated with methodology pitfalls like reverse causation, i.e. low body weight may be an index for the presence of chronic diseases like cancer, malnutrition, infectious disease, smoking duration and intensity, which in turn increase mortality.[264] For example, in a National Health and Nutrition Examination Survey (NHANES) analysis, the obesity paradox was present among persons with abnormality in blood glucose levels, but was absent in the subgroup of never-smokers.[265] To identify potential reverse causation in the analyses, I performed subgroup analyses in patients with and without diabetes mellitus, current smoking, and patients without cancer diagnosis at the time of incident stroke. In this cohort, diabetes mellitus was less prevalent while current smoking and cancer were more prevalent in underweight patients. The stroke-obesity paradox was present irrespective of diabetes mellitus (see Appendix F.6.5), smoking status (see Appendix F.6.6), or cancer at time of incident stroke (see Table 6.4). Although these findings do not support the explanation of reverse causation, it may still be possible that this might have occurred by other chronic illnesses that I did not consider in these analyses.

Another suggested explanation for the stroke-obesity paradox was residual confounding.[264] In this study, the results were adjusted for many prospectively registered patient characteristics like age, sex, socioeconomic status, comorbidities, and concurrent medication. I cannot exclude the possibility that additional unmeasured confounding bias might have been introduced, for example, comorbidities that are associated with cardiovascular outcomes might have not been equally distributed among BMI strata. However, key comorbidities of cardiovascular risk are featured within the metabolic syndrome comprising hypertension, diabetes, hyperlipidaemia all well associated with excessive body weight. Those factors were well included in the multivariable-adjusted assessments and – in accordance with common knowledge – a higher, not lower, prevalence of such comorbidities with higher body weight was observed in this study. Hence a higher risk profile of relevant cardiovascular risk factors may be concluded for patients with higher BMI. The main strengths of this analysis can be seen in the large size of this prospective population-based cohort, the long duration of follow-up exceeding a decade, and a large number of outcome events. Moreover, to minimize the risk of bias due to residual confounding, the results were adjusted for a wide range of comorbidities and clinical covariates. Also, to identify potential reverse causal pathways, I performed subgroup analyses according to sex, diabetes, current smoking habit, and cancer (excluding those with a diagnosis) at the time of incident stroke.

A limitation of the study was that BMI was the only marker of obesity that was analysed, as there were no available data about other anthropometric markers of obesity like waist-hip ratio or waist circumference. Waist-to-hip ratio or waist circumference is a more precise measurement of obesity. In a study by Janseen et al., the obesity paradox was non-existent when BMI was replaced by waist-to-hip ratio.[266] Given that BMI is an imperfect marker of obesity, it would be interesting to see in other cohorts whether the obesity-paradox remains present

when other markers of obesity are analysed. Moreover, combined models showed that within BMI groups, waist circumference can further stratify cardiovascular risk.[267,268] Recently, an analysis in the ORIGIN dataset identified weight loss as an independent risk factor for higher mortality compared to no weight loss.[269] The obesity paradox has also been considered to be the result of potential survival bias.[270] The possibility of selection bias due to a survival bias cannot be ruled out in this study. It is important to note that the conclusions of this analysis as well as previous reports of the stroke-obesity paradox, should only be viewed as a putative association and should not be perceived as proof of causality. Therefore, no recommendations about weight management after stroke should be based on these conclusions. Ongoing randomized controlled trials might provide further evidence to guide weight management recommendations in stroke survivors. Semaglutide was recently associated with a sustained, clinically relevant reduction in body weight[271] and is currently assessed for the reduction of cardiovascular events in patients with overweight or obesity and prior cardiovascular disease including stroke.[272]

6.6 Conclusion

In this prospective population-based cohort study of 30,702 patients with incident stroke followed for a median duration of 12.9 years, overweight or obesity was associated with a more favourable prognosis for subsequent MACE, PVD, and mortality, irrespective of sex, diabetes mellitus, smoking or cancer at the time of incident stroke.

Summary

This chapter assessed the relationship between BMI and subsequent MACE outcomes in patients with incident stroke. The next chapter uses a novel cluster analysis (a hypothesis-free unsupervised machine learning data-driven approach) to classify patients with incident stroke into phenotypic clusters and evaluate the differential burden of subsequent cardiovascular morbidity and mortality outcomes.

Chapter 7

A population-based study exploring phenotypic clusters and clinical outcomes in stroke using an unsupervised machine learning approach

The range of studies in previous chapters have explored existing evidence on the prediction of outcomes after incident stroke; and have reported variations in subsequent cardiovascular morbidity and mortality outcomes after incident stroke associated with risk factors including age, sex, socioeconomic status, nature/type of incident stroke, and BMI. This chapter uses a novel cluster analysis approach for mixed data (i.e., both categorical and continuous) to classify patients with incident stroke into phenotypic clusters and evaluates the differential burden of subsequent major adverse cardiovascular morbidity and mortality outcomes.

A manuscript based on this study is under peer-review with the journal *BMC Medical Informatics and Decision Making*:

Akyea, R.K., Ntaios, G., Kontopantelis, E., Georgiopoulos, G., Soria, D., Asselbergs, F.W., Kai, J., Weng, S.F., Qureshi, N. A population-based study exploring phenotypic clusters and clinical outcomes in stroke using unsupervised machine learning approach.

7.1 Abstract

Background: Individuals developing stroke have varying clinical characteristics, demographic, and biochemical profiles. This heterogeneity in phenotypic characteristics comprising sociodemographic, biological, and comorbidity profiles can impact cardiovascular disease (CVD) morbidity and mortality outcomes. Cluster analysis, a hypothesis-free unsupervised machine learning approach, has been widely used to put heterogeneous populations into relatively homogenous clusters (subgroups) with similar characteristics.

Objective: This study used a novel cluster analysis approach to stratify individuals with incident stroke into phenotypic clusters and evaluated the differential burden of cardiovascular morbidity and mortality outcomes.

Methods: Linked clinical data from primary care, hospitalisations, social deprivation, and death records in the UK, were used to cluster 48,114 adult patients based on their demographic, biochemical, comorbidities, and prescribed medication profiles at the time of incident stroke. A data-driven cluster analysis (kamila algorithm) was used. Cox proportional hazards regression was used to estimate hazard ratios (HRs) for subsequent adverse outcomes, for each of the generated clusters. Subsequent outcomes included CHD, recurrent stroke, PVD, heart failure, CVD-related and all-cause mortality.

Results: Four distinct phenotypic cohorts with varying underlying clinical characteristics were identified in patients with incident stroke. Cluster 1 (n=5,201, 10.8%) was a cohort with high prevalence of CHD-related risk factors and prescribed medications; cluster 2 (n=18,655, 38.8%) a cohort with low prevalence of multiple long-term conditions (MLTC); cluster 3 (n=10,244, 21.3%) a cohort with high prevalence of MLTC; and cluster 4 (n=14,014, 29.1%), the oldest population cohort and predominantly female. Compared to cluster 1, the risk of composite recurrent stroke or CVD-related mortality outcome was higher

in the other 3 clusters (cluster 2: hazard ratio [HR], 1.07; 95% CI, 1.02-1.12; cluster 3: HR, 1.20; 95% CI, 1.14-1.26; and cluster 4: HR, 1.44; 95% CI: 1.37-1.50). Similar trends in risk were observed for composite recurrent stroke and all-cause mortality outcome, and subsequent recurrent stroke outcome. However, results were not consistent for subsequent risk in CHD, PVD, heart failure, CVD-related mortality, and all-cause mortality. The risk of subsequent heart failure, CVD-related and all-cause mortality were significantly decreased for patients in cluster 2 while patients in clusters 3 and 4 had a significantly increased risk when compared to cluster 1.

Conclusions: This proof of principle study, demonstrates how a heterogenous population of patients with incident stroke can be stratified into four relatively homogenous phenotypes with differential risk of subsequent cardiovascular morbidity and mortality outcomes. This offers an opportunity to revisit the stratification of patients with incident stroke and highlights the potential to target modifiable characteristics in clusters for more targeted preventive intervention.

7.2 Introduction

Patients at the time of incident stroke have varied clinical characteristics, demographics, socioeconomic, biochemical, comorbidity, and prescribed medication profiles. This heterogeneity in characteristics at the time of incident stroke impacts on cardiovascular morbidity and mortality outcomes.[273] Phenotyping (subgrouping) people after incident stroke, in terms of the risk of various cardiovascular outcomes, could provide individuals with the poorest prognosis better care. Intensive secondary prevention strategies including the use of novel medications such as proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitors and colchicine in patients at very high risk of adverse cardiovascular morbidity and mortality outcomes.

Cluster analysis, a hypothesis-free unsupervised machine learning data-driven approach, has been widely used to analyse clinical data to identify new phenotypic subgroups of complex and heterogeneous diseases including obstructive sleep apnoea,[274] asthma,[275,276] chronic obstructive pulmonary disease, chronic heart failure,[277] dilated cardiomyopathy,[278] sepsis,[279] Parkinson's disease,[280] breast cancer,[281] and diabetes.[6] This approach does not include outcome data and may be less biased in its results, especially when using retrospectively collected data. Clustering of clinical data may, therefore, help identify subgroups of patients with incident stroke and generate new hypotheses. Efforts to determine such phenotypic groups in patients with incident stroke remain limited.

Using a large population-based cohort of adult patients with incident stroke, the objectives of this study are: (i) to identify patterns in linked primary and secondary clinical data and cluster patients based on phenotypic characteristics; (ii) to assess the association between phenotypic clusters and subsequent cardiovascular morbidity and mortality outcomes (i.e. composite of recurrent stroke or CVD-related mortality; composite of recurrent stroke or all-cause

mortality; CHD, recurrent stroke, PVD, heart failure, CVD-related mortality, and all-cause mortality).

7.3 Methods

Study design and data source

This prospective population-based cohort study used the UK CPRD GOLD database,[87] linked to HES APC,[208] national mortality data,[95] and social deprivation data.[97] The databases have been previously described in Chapter 2 (Section 2.2).

Study population

The study cohort of patients with the first record of non-fatal stroke in either CPRD GOLD or HES APC between 1 January 1998 and 31 December 2017 has been previously described in Chapter 2 (Section 2.3.1).

Outcomes

The primary outcome was a composite of either recurrent stroke or CVD-related mortality events recorded after incident stroke from across the linked data sources (CPRD, HES or ONS registry). The secondary outcomes included: CHD, recurrent stroke, PVD, heart failure, CVD-related mortality, all-cause mortality, and the composite of either recurrent stroke or all-cause mortality.

Subsequent outcomes within 30 days were considered to be representing or relating to the incident stroke event.[282] Analyses were, therefore, restricted to patients with subsequent outcomes occurring beyond 30 days after incident stroke.

Potential candidate variables for phenotyping

Based on availability in the electronic health records and established association with CVD, 336 candidate variables were selected – outlined in Chapter 2 (Section 2.3.2.3). These included demographic data, vital signs, biochemical parameters,

comorbid conditions, and prescribed medications (Appendix G.7 Table 1). For vital signs and biochemical test results, the most recent values/records within 24 months before incident stroke were extracted. A prescription within 12 months before the incident stroke was considered as a medication prescribed. All comorbid conditions were defined based on the latest record of a comorbid condition any time before the incident stroke.

Data processing

The variable distributions and missingness were first assessed. As described in Chapter 2 (Section 2.3.4), multiple imputation by chained equations was used to generate imputed datasets to account for missing data (Appendix G.7 Figure 1, Appendix G.7 Table 1).[108–110] The imputed datasets were pooled into a single dataset using Rubin's rules.[111] A high number of dimensions from a dataset with many variables/features is associated with a loss of meaningful differentiation between similar and dissimilar individuals – 'curse of dimensionality.[283] To improve the cluster analysis process and performance, feature selection was done to eliminate redundant variables. Feature selection was based on two (2) widely used data-driven feature selection methods (Boruta[284] and Least Absolute Shrinkage and Selection Operator (Lasso) regression[285] – Appendix G.7 Figure 2) and clinical expert consensus. An expert group of 4 clinicians from both primary care (General Practitioners – NQ, JK) and secondary care (Stroke Medicine Consultant/Specialist – GN, GG) settings were independently consulted to attain consensus on which variables to select for the cluster analysis. Clinical expert consensus was defined as a 75% (3 out of 4) agreement among the clinical experts on each variable. 49 variables were rated important by the clinical experts and at least 1 of the 2 data-driven methods – Table 7.1.

Table 7.1 Overview of all variables and the in- or exclusion at the various data processing steps

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Sex	Demographics	Men: 31,389 (45.7) Women: 37,253 (54.3)	X	X	X	X	**X**
Age at incident stroke, years	Demographics	Mean: 73.3 (SD: 13.9)	X	X	X	X	**X**
Incident stroke sub-type		Haemorrhagic: 6,535 (9.5) Ischaemic: 25,556 (37.2) Stroke NOS: 36,551 (53.2)	X	X	X	X	**X**
Year of incident stroke		1998 – 2017	X	X			
Ethnicity	Demographics	Asian: 891 (1.3) Black: 557 (0.8) Mixed: 102 (0.1) Other: 480 (0.7) White: 60,937 (88.8) Unknown: 5,675 (8.3)	X	X	X	X	**X**
Index of multiple deprivations	Socio-economic status	1: 14,740 (21.5) 2: 15,289 (22.3) 3: 14,828 (21.6) 4: 12,613 (18.4) 5: 11,056 (16.1) Unknown: 116 (0.2)			X		
Smoking status	Lifestyle	Never: 26,229 (38.2) Ex: 16,080 (23.4) Current: 12,102 (17.6) Unknown: 14,231 (20.7)	X		X	X	**X**
Alcohol status	Lifestyle	Yes: 15,822 (23.1) No: 5,248 (7.6) Ex: 1,177 (1.7) Unknown: 46,395 (67.6)			X		
Physical measurements							
Body mass index		26.4 (25.1 – 27.9)		X	X	X	**X**
Diastolic blood pressure	Vital sign	80 (74 – 84)		X	X	X	**X**
Systolic blood pressure	Vital sign	140 (130 – 149)	X	X	X	X	**X**

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Height		1.7 (1.6 – 1.7)					
Pulse	Vital sign	76 (73 – 79)		X	X	X	X
Weight	Biochemical test	73.9 (68.0 – 79.5)		X	X	X	
Biochemical tests							
Alanine aminotransferase	Biochemical test	22.35 (19.0 – 26.08)		X			
Albumin level	Biochemical test	40.65 (39.0 – 42.0)		X			
Alkaline phosphatase	Biochemical test	91.0 (77.8 – 103.0)	X	X			
Bilirubin level	Biochemical test	10.55 (9.0 – 12.0)		X			
Calcium level (adjusted)	Biochemical test	2.34 (2.31 – 2.36)		X			
Calcium level	Biochemical test	2.34 (2.31 – 2.37)		X			
Creatinine level	Biochemical test	90.48 (80.0 – 100.0)		X			
C-reactive protein	Biochemical test	10.0 (6.45 – 15.0)		X	X	X	X
Eosinophil level	Biochemical test	0.26 (0.16 – 0.38)		X			
Erythrocyte sedimentation rate	Biochemical test	18.0 (12.9 – 23.03)		X			
Gamma glutamyl transpeptidase	Biochemical test	43.13 (32.93 – 57.83)		X			
Glomerular filtration rate	Biochemical test	67.04 (62.23 – 71.90)		X	X	X	X
Haemoglobin level	Biochemical test	13.53 (12.9 – 14.2)	X	X	X	X	X
Glycated haemoglobin (hba1c) level	Biochemical test	50.0 (46.79 – 53.46)	X	X	X	X	X
HDL/LDL ratio	Biochemical test	3.65 (3.22 – 4.10)	X	X	X	X	
High-density lipoprotein (HDL) cholesterol	Biochemical test	1.47 (1.30 – 1.63)		X	X	X	X
Low-density lipoprotein (LDL) cholesterol	Biochemical test	2.97 (2.66 – 3.26)	X	X	X	X	X
Lymphocyte count	Biochemical test	2.40 (1.71 – 3.33)		X			
Neutrophil count	Biochemical test	4.74 (4.11 – 5.58)		X			
Platelet count	Biochemical test	248.0 (221.3 – 275.5)		X			
Potassium level	Biochemical test	4.4 (4.2 – 4.5)		X			
Sodium level	Biochemical test	139 (138 – 141)		X			
Thyroid-stimulating hormone level	Biochemical test	20.7 (1.79 – 2.32)	X				
Total cholesterol level	Biochemical test	5.09 (4.70 – 5.45)		X	X	X	
Triglyceride level	Biochemical test	1.43 (1.21 – 1.67)		X	X	X	X
Urea	Biochemical test	6.3 (5.4 – 7.1)		X			

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Comorbid conditions							
Benign neoplasm – brain	Benign neoplasm	303 (0.4)		X			
Benign neoplasm - colon	Benign neoplasm	1,093 (1.6)					
Benign neoplasm - ovary	Benign neoplasm	493 (0.7)					
Benign neoplasm - stomach	Benign neoplasm	165 (0.2)					
Benign neoplasm - uterus	Benign neoplasm	149 (0.2)					
Haemangioma	Benign neoplasm	520 (0.8)					
Leiomyoma	Benign neoplasm	573 (0.8)					
Cancer *(composite)*	Cancers	11,111 (16.2)	X	X	X	X	**X**
Hodgkin Lymphoma	Cancers	34 (0.0)					
Leukaemia	Cancers	231 (0.3)	X	X			
Metastatic tumour	Cancers	333 (0.5)	X	X			
Monoclonal gammopathy of uncertain significance	Cancers	142 (0.2)					
Myelodysplastic syndrome	Cancers	139 (0.2)					
Non-Hodgkin Lymphoma	Cancers	294 (0.4)					
Non-metastatic cancer	Cancers	5,955 (8.7)					
Plasma cell malignancy	Cancers	120 (0.2)					
Polycythaemia vera	Cancers	145 (0.2)					
Primary malignancy – biliary	Cancers	20 (0.0)					
Primary malignancy – bladder	Cancers	393 (0.6)					
Primary malignancy – bone	Cancers	16 (0.0)					
Primary malignancy – bowel	Cancers	880 (1.3)					
Primary malignancy – brain	Cancers	114 (0.2)	X	X			
Primary malignancy – breast	Cancers	1,599 (2.3)					
Primary malignancy – cervical	Cancers	78 (0.1)					
Primary malignancy – kidney	Cancers	101 (0.1)					
Primary malignancy – liver	Cancers	10 (0.0)					
Primary malignancy – lung	Cancers	342 (0.5)	X	X			
Primary malignancy – melanoma	Cancers	544 (0.8)					

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Primary malignancy – oesophageal	Cancers	107 (0.2)	X				
Primary malignancy – oropharyngeal	Cancers	114 (0.2)					
Primary malignancy – other	Cancers	367 (0.5)					
Primary malignancy – ovarian	Cancers	110 (0.2)					
Primary malignancy – pancreas	Cancers	41 (0.1)					
Primary malignancy – prostate	Cancers	1,135 (1.6)					
Primary malignancy – skin	Cancers	4,283 (6.2)		X			
Primary malignancy – stomach	Cancers	67 (0.1)					
Primary malignancy – testis	Cancers	26 (0.0)					
Primary malignancy – thyroid	Cancers	26 (0.0)					
Primary malignancy – uterus	Cancers	152 (0.2)					
Secondary malignancy – bone	Cancers	59 (0.1)					
Secondary malignancy – brain	Cancers	36 (0.0)	X	X			
Secondary malignancy – liver	Cancers	66 (0.1)		X			
Secondary malignancy – lung	Cancers	27 (0.0)					
Secondary malignancy – lymph nodes	Cancers	30 (0.0)					
Secondary malignancy – others	Cancers	281 (0.4)	X	X			
Abdominal aortic aneurysm	Diseases – circulatory system	457 (0.7)			X		
Arrythmia	Diseases – circulatory system	6,983 (10.2)	X	X	X	X	**X**
Atrial fibrillation	Diseases – circulatory system	6,453 (9.4)	X	X	X	X	
Atrioventricular block, first degree	Diseases – circulatory system	39 (0.1)					
Atrioventricular block, second degree	Diseases – circulatory system	14 (0.0)					
Atrioventricular block, third-degree	Diseases – circulatory system	35 (0.0)					
Cardiomyopathy - other	Diseases – circulatory system	68 (0.1)			X		
Dilated cardiomyopathy	Diseases – circulatory system	26 (0.0)			X		
Family history of cardiovascular disease	Diseases – circulatory system	12,299 (17.9)			X		
Family history of coronary heart disease	Diseases – circulatory system	8,575 (12.5)			X		
Hypertension	Diseases – circulatory system	31,844 (46.4)	X		X	X	**X**
Hypertrophic cardiomyopathy	Diseases – circulatory system	38 (0.1)			X		
Left bundle branch block	Diseases – circulatory system	88 (0.1)			X		

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Multiple valve disorder	Diseases – circulatory system	104 (0.1)			X		
Non-rheumatic aortic valve disorder	Diseases – circulatory system	834 (1.2)	X		X	X	**X**
Non-rheumatic mitral valve disorder	Diseases – circulatory system	618 (0.9)			X		
Pericardial effusion	Diseases – circulatory system	35 (0.0)					
Primary pulmonary hypertension	Diseases – circulatory system	52 (0.1)			X		
Raynaud's disease	Diseases – circulatory system	752 (1.1)					
Rheumatic valve disorder	Diseases – circulatory system	141 (0.2)			X		
Right bundle branch block	Diseases – circulatory system	130 (0.2)			X		
Sick sinus syndrome	Diseases – circulatory system	74 (0.1)					
Subarachnoid haemorrhage	Diseases – circulatory system	477 (0.7)	X	X			
Subdural haematoma	Diseases – circulatory system	114 (0.2)		X			
Supraventricular tachycardia	Diseases – circulatory system	629 (0.9)			X		
Transient ischaemic attack	Diseases – circulatory system	14,068 (20.5)		X	X	X	**X**
Venous thrombolism (excluding PR)	Diseases – circulatory system	1,789 (2.6)					
Ventricular tachycardia	Diseases – circulatory system	64 (0.1)			X		
Alcoholic liver disease	Diseases – digestive system	260 (0.4)			X		
Autoimmune liver disease	Diseases – digestive system	55 (0.1)					
Barrett's Oesophagus	Diseases – digestive system	505 (0.7)					
Cholangitis	Diseases – digestive system	101 (0.1)					
Cholecystitis	Diseases – digestive system	745 (1.1)					
Cholelithiasis	Diseases – digestive system	2,183 (3.2)					
Cirrhosis	Diseases – digestive system	334 (0.5)					
Coeliac disease	Diseases – digestive system	221 (0.3)					
Crohn's disease	Diseases – digestive system	209 (0.3)					
Diverticular disease	Diseases – digestive system	4,851 (7.1)					
Fatty liver	Diseases – digestive system	56 (0.1)			X		
Gastritis and duodenitis	Diseases – digestive system	3,680 (5.4)		X			
Gastroesophageal reflux disease	Diseases – digestive system	6,339 (9.2)					
Irritable bowel syndrome	Diseases – digestive system	3,264 (4.8)					
Liver failure	Diseases – digestive system	43 (0.1)					

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Mild liver disease	Diseases – digestive system	212 (0.3)					
Moderate-severe liver disease	Diseases – digestive system	316 (0.5)		X			
Pancreatitis	Diseases – digestive system	449 (0.6)					
Peptic ulcer disease	Diseases – digestive system	2,633 (3.8)					
Peritonitis	Diseases – digestive system	308 (0.4)					
Portal hypertension	Diseases – digestive system	32 (0.0)					
Ulcerative colitis	Diseases – digestive system	423 (0.6)					
Hearing loss	Diseases – Ear	10,587 (15.4)					
Meniere's disease	Diseases – Ear	524 (0.8)					
Otitis media	Diseases – Ear	3,616 (5.3)					
Tinnitus	Diseases – Ear	3,023 (4.4)					
Cystic fibrosis	Diseases – Endocrine system	23 (0.0)					
Diabetes mellitus	Diseases – Endocrine system	7,978 (11.6)		X	X	X	**X**
Diabetes mellitus, Type 1	Diseases – Endocrine system	577 (0.8)			X		
Diabetes mellitus, Type 2	Diseases – Endocrine system	6,578 (9.6)		X	X	X	
Diabetes mellitus, with complications	Diseases – Endocrine system	1,404 (2.0)			X		
Diabetes mellitus, with no complications	Diseases – Endocrine system	7,946 (11.6)		X	X	X	
Dyslipidaemia	Diseases – Endocrine system	6,560 (9.6)		X	X	X	**X**
Family history of hyperlipidaemia	Diseases – Endocrine system	86 (0.1)			X		
Hyperparathyroidism	Diseases – Endocrine system	224 (0.3)					
Hypoglycaemia-causing disorders	Diseases – Endocrine system	234 (0.3)					
Hypothyroidism	Diseases – Endocrine system	4,869 (7.1)					
Obesity	Diseases – Endocrine system	3,096 (4.5)			X		
Polycystic ovarian syndrome	Diseases – Endocrine system	24 (0.0)			X		
Thyroid disease *(hypo or hyperthyroidism)*	Diseases – Endocrine system	5,708 (8.3)					
Anterior uveitis	Diseases – Eye	694 (1.0)					
Blindness	Diseases – Eye	2,696 (3.9)					
Cataract	Diseases – Eye	10,776 (15.7)	X	X			
Diabetic ophthalmic complications	Diseases – Eye	1,986 (2.9)	X		X	X	**X**
Glaucoma	Diseases – Eye	3,293 (4.8)					

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Keratitis	Diseases – Eye	388 (0.6)					
Macular degeneration	Diseases – Eye	2,307 (3.4)					
Posterior uveitis	Diseases – Eye	47 (0.1)					
Retinal detachment	Diseases – Eye	522 (0.8)					
Retinal vascular occlusion	Diseases – Eye	959 (1.4)					
Scleritis	Diseases – Eye	354 (0.5)					
Acute kidney injury	Diseases – genitourinary system	314 (0.5)		X	X	X	**X**
Benign prostatic hyperplasia	Diseases – genitourinary system	4,270 (6.2)					
Chronic kidney disease	Diseases – genitourinary system	7,232 (10.5)		X	X	X	
End-stage renal disease	Diseases – genitourinary system	239 (0.4)			X		
Erectile dysfunction	Diseases – genitourinary system	3,735 (5.4)			X		
Female infertility	Diseases – genitourinary system	197 (0.3)					
Glomerulonephritis	Diseases – genitourinary system	144 (0.2)			X		
Male infertility	Diseases – genitourinary system	187 (0.3)					
Neuropathic bladder	Diseases – genitourinary system	1,012 (1.5)					
Obstructive and reflux uropathy	Diseases – genitourinary system	289 (0.4)					
Proteinuria	Diseases – genitourinary system	866 (1.3)			X		
Renal disease	Diseases – genitourinary system	8,108 (11.8)		X	X	X	**X**
Urinary incontinence	Diseases – genitourinary system	4,913 (7.2)	X	X			
Urolithiasis	Diseases – genitourinary system	1,599 (2.3)					
Allergic and chronic rhinitis	Diseases – respiratory system	7,024 (10.2)					
Asbestosis	Diseases – respiratory system	117 (0.2)					
Asthma	Diseases – respiratory system	6,771 (9.9)	X				
Bronchiectasis	Diseases – respiratory system	425 (0.6)					
Chronic obstructive pulmonary disease	Diseases – respiratory system	3,932 (5.7)		X			
Chronic sinusitis	Diseases – respiratory system	4,357 (6.3)					
Pleural effusion	Diseases – respiratory system	305 (0.4)					
Pleural plaque	Diseases – respiratory system	86 (0.1)					
Pneumothorax	Diseases – respiratory system	228 (0.3)					
Pulmonary collapse	Diseases – respiratory system	69 (0.1)					

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Pulmonary fibrosis	Diseases – respiratory system	166 (0.2)	X				
Respiratory failure	Diseases – respiratory system	28 (0.0)					
Sleep apnoea	Diseases – respiratory system	335 (0.5)			X		
Agranulocytosis	Haem. / Immunological conditions	272 (0.4)					
Anaemia – other	Haem. / Immunological conditions	4,567 (6.6)		X			
Aplastic anaemia	Haem. / Immunological conditions	53 (0.1)	X				
Folate deficiency anaemia	Haem. / Immunological conditions	378 (0.5)					
Hypersplenism	Haem. / Immunological conditions	45 (0.1)					
Hyposplenism	Haem. / Immunological conditions	151 (0.2)					
Immunodeficiency	Haem. / Immunological conditions	17 (0.0)					
Iron deficiency anaemia	Haem. / Immunological conditions	3,023 (4.4)					
Other haemolytic anaemia	Haem. / Immunological conditions	81 (0.1)					
Primary thrombocytopaenia	Haem. / Immunological conditions	83 (0.1)					
Sarcoidosis	Haem. / Immunological conditions	139 (0.2)					
Secondary polycythaemia	Haem. / Immunological conditions	83 (0.1)					
Secondary thrombocytopaenia	Haem. / Immunological conditions	265 (0.4)					
Sickle cell trait	Haem. / Immunological conditions	26 (0.0)					
Thalassaemia	Haem. / Immunological conditions	30 (0.0)					
Thalassaemia trait	Haem. / Immunological conditions	42 (0.1)					
Thrombophilia	Haem. / Immunological conditions	141 (0.2)			X		
Vitamin B12 deficiency anaemia	Haem. / Immunological conditions	1,682 (2.4)					
Chronic viral hepatitis	Infectious diseases	128 (0.2)					
HIV	Infectious diseases	24 (0.0)					
Rheumatic fever	Infectious diseases	245 (0.4)					
Tuberculosis	Infectious diseases	575 (0.8)					
Alcohol misuse	Mental health disorders	1,903 (2.8)	X	X	X	X	**X**
Anxiety	Mental health disorders	8,782 (12.8)					
Autism	Mental health disorders	18 (0.0)					
Bipolar affective disorder	Mental health disorders	352 (0.5)			X		
Conduct disorder	Mental health disorders	66 (0.1)					

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Delirium	Mental health disorders	567 (0.8)					
Dementia	Mental health disorders	3,532 (5.1)	X	X	X	X	**X**
Depression	Mental health disorders	12,597 (18.4)		X	X	X	**X**
Eating disorders	Mental health disorders	62 (0.1)					
Hyperkinetic disorders	Mental health disorders	28 (0.0)					
Intellectual disability	Mental health disorders	264 (0.4)					
Insomnia	Mental health disorders	6,902 (10.1)					
Obsessive-compulsive disorder	Mental health disorders	155 (0.2)					
Personality disorder	Mental health disorders	371 (0.5)					
Schizophrenia	Mental health disorders	626 (0.9)		X			
Self-harm	Mental health disorders	1,495 (2.2)		X			
Severe mental illness	Mental health disorders	955 (1.4)	X	X	X	X	**X**
Substance misuse	Mental health disorders	701 (1.0)			X		
Ankylosing spondylitis	Musculoskeletal conditions	102 (0.1)					
Back pain	Musculoskeletal conditions	24,933 (36.3)					
Carpal tunnel syndrome	Musculoskeletal conditions	3,156 (4.6)					
Collapsed vertebra	Musculoskeletal conditions	383 (0.6)					
Connective tissue disease	Musculoskeletal conditions	3,245 (4.7)					
Enthesopathy and synovial disorder	Musculoskeletal conditions	14,198 (20.7)					
Fibromatosis	Musculoskeletal conditions	1,367 (2.0)					
Giant cell arteritis	Musculoskeletal conditions	414 (0.6)					
Gout	Musculoskeletal conditions	3,837 (5.6)			X		
Intervertebral disc disorder	Musculoskeletal conditions	1,699 (2.5)					
Lupus erythematosus	Musculoskeletal conditions	152 (0.2)			X		
Osteoarthritis	Musculoskeletal conditions	16,995 (24.8)	X				
Osteoporosis	Musculoskeletal conditions	4,434 (6.5)		X			
Polymyalgia rheumatica	Musculoskeletal conditions	1,842 (2.7)					
Psoriatic arthritis	Musculoskeletal conditions	165 (0.2)					
Reactive arthritis	Musculoskeletal conditions	34 (0.0)					
Rheumatoid arthritis	Musculoskeletal conditions	1,317 (1.9)			X		

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Scleroderma	Musculoskeletal conditions	32 (0.0)					
Scoliosis	Musculoskeletal conditions	448 (0.6)					
Sjogren syndrome	Musculoskeletal conditions	123 (0.2)					
Spinal stenosis	Musculoskeletal conditions	597 (0.9)					
Spondylolisthesis	Musculoskeletal conditions	255 (0.4)					
Spondylosis	Musculoskeletal conditions	6,171 (9.0)					
Autonomic neuropathy	Neurological conditions	210 (0.3)					
Bell's palsy	Neurological conditions	695 (1.0)					
Cerebral palsy	Neurological conditions	64 (0.1)					
Chronic fatigue syndrome	Neurological conditions	1,026 (1.5)					
Diabetic neuropathy	Neurological conditions	409 (0.6)			X		
Epilepsy	Neurological conditions	1,876 (2.7)	X	X			
Essential tremor	Neurological conditions	331 (0.5)					
Hemiplegia	Neurological conditions	342 (0.5)	X				
Migraine	Neurological conditions	3,610 (5.3)					
Motor neurone disease	Neurological conditions	28 (0.0)					
Multiple sclerosis	Neurological conditions	214 (0.3)					
Myasthenia gravis	Neurological conditions	52 (0.1)					
Parkinson's disease	Neurological conditions	959 (1.4)	X	X			
Peripheral neuropathy	Neurological conditions	1,833 (2.7)					
Trigeminal neuralgia	Neurological conditions	661 (1.0)					
Congenital septal defect	Perinatal conditions	110 (0.2)					
Acne	Skin conditions	693 (1.0)	X	X			
Actinic keratosis	Skin conditions	3,400 (5.0)					
Alopecia areata	Skin conditions	148 (0.2)					
Dermatitis	Skin conditions	13,950 (20.3)					
Hidradenitis supprativa	Skin conditions	87 (0.1)					
Lichen planus	Skin conditions	536 (0.8)					
Pilonidal cyst/sinus	Skin conditions	223 (0.3)					
Psoriasis	Skin conditions	2,588 (3.8)					

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Rosacea	Skin conditions	1,661 (2.4)					
Seborrheic dermatitis	Skin conditions	3,268 (4.8)					
Urticaria	Skin conditions	2,388 (3.5)					
Vitiligo	Skin conditions	130 (0.2)					
Prescribed medications							
Acarbose	Prescribed medication	118 (0.2)					
Angiotensin-converting enzyme inhibitor	Prescribed medication	20,145 (29.3)			X		
Alpha-blocker	Prescribed medication	4,267 (6.2)					
Antihypertensive	Prescribed medication	33,347 (48.6)		X	X	X	**X**
Antiarrhythmic	Prescribed medication	3,152 (4.6)			X		
Anticoagulant	Prescribed medication	4,050 (5.9)		X	X	X	**X**
Antidepressant	Prescribed medication	6,368 (9.3)	X	X	X	X	**X**
Antidiabetic	Prescribed medication	15,474 (22.5)		X	X	X	**X**
Antiepileptic	Prescribed medication	5,679 (8.3)	X	X			
Antiplatelet	Prescribed medication	25,676 (37.4)			X		
Anxiolytic	Prescribed medication	7,709 (11.2)	X				
Beta blocker	Prescribed medication	15,693 (22.9)	X	X	X	X	
Bile acid sequestrant	Prescribed medication	106 (0.1)					
Calcium channel blocker	Prescribed medication	16,493 (24.0)			X		
Centrally acting antihypertensive	Prescribed medication	699 (1.0)			X		
Corticosteroid	Prescribed medication	6,715 (9.8)			X		
Diuretic	Prescribed medication	24,114 (35.1)		X	X	X	**X**
DPP-4 inhibitors (Gliptins)	Prescribed medication	319 (0.5)			X		
Fibrates	Prescribed medication	210 (0.3)			X		
Glinide	Prescribed medication	51 (0.1)					
Glucagon-like peptide-1 (GLP-1)	Prescribed medication	70 (0.1)			X		
Hormone replacement therapy	Prescribed medication	1,024 (1.5)					
Immunosuppressant	Prescribed medication	6,587 (9.6)			X		
Inotrope	Prescribed medication	3,347 (4.9)	X	X	X	X	**X**
Loop diuretic	Prescribed medication	9,518 (13.9)		X	X	X	**X**

Variables	Domain	Prevalence, n (%)	LASSO	Boruta	Clinical experts	Selected variables	Cluster analysis
Metformin	Prescribed medication	4,524 (6.6)		X	X		
Nicotinic acid	Prescribed medication	11 (0.0)					
Nitrates	Prescribed medication	1,571 (2.3)			X		
Non-steroidal anti-inflammatory drugs	Prescribed medication	17,579 (25.6)					
Opioid	Prescribed medication	26,910 (39.2)	X				
Oral contraception	Prescribed medication	231 (0.3)					
Peripheral vasodilator	Prescribed medication	180 (0.3)			X		
Proton pump inhibitor	Prescribed medication	18,515 (27.0)	X	X			
RAAS inhibitor	Prescribed medication	20,142 (29.3)			X		
Sodium-glucose co-transporter-2 inhibitors	Prescribed medication	21 (0.0)			X		
Statin potency	Prescribed medication	Low: 2,440 (3.5) Moderate: 12,511 (18.2) High: 2,917 (4.2)		X	X	X	**X**
Sulfonylureas	Prescribed medication	3,221 (4.7)			X		
Thiazide diuretic	Prescribed medication	16,505 (24.1)	X	X	X	X	**X**
Thiazolidinediones	Prescribed medication	597 (0.9)			X		
Vasodilator	Prescribed medication	283 (0.4)			X		
Warfarin	Prescribed medication	3,696 (5.4)		X	X	X	

After evaluating correlation among the 49 selected variables using mixedCor and Lares functions in R for mixed-type data (Figures 7.1 & 7.2), 10 highly correlated variables were excluded based on clinical judgement/importance – diagnosis of atrial fibrillation, chronic kidney disease, diabetes mellitus (with no complications), diabetes mellitus (type-2), HDL/LDL ratio, total cholesterol, weight, prescription of beta-blocker, metformin, and warfarin. The remaining 39 variables, Box 7.1, were used for the cluster analysis.

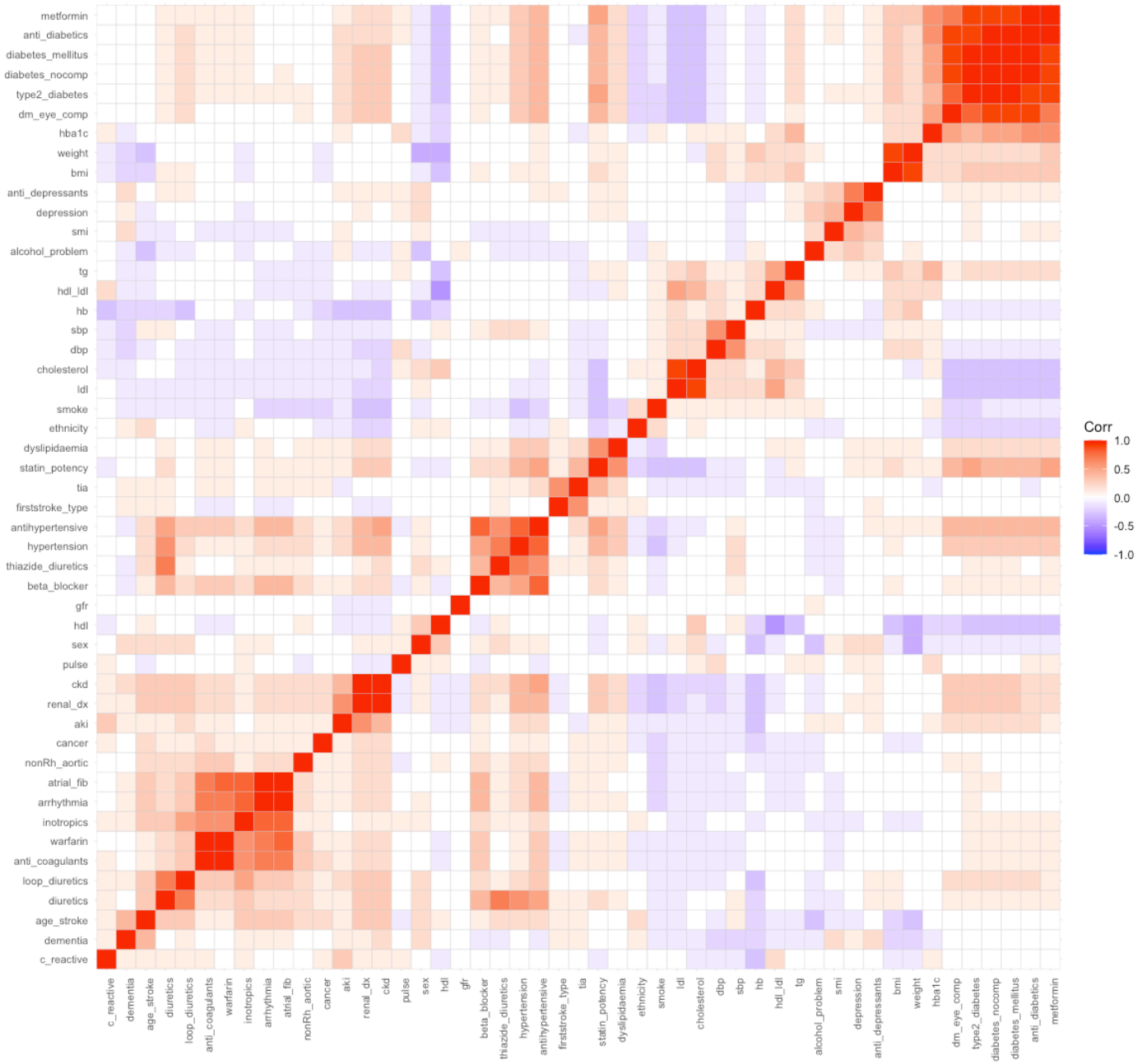


Figure 7.1 **Plot of the correlation matrix of 49 selected variables**

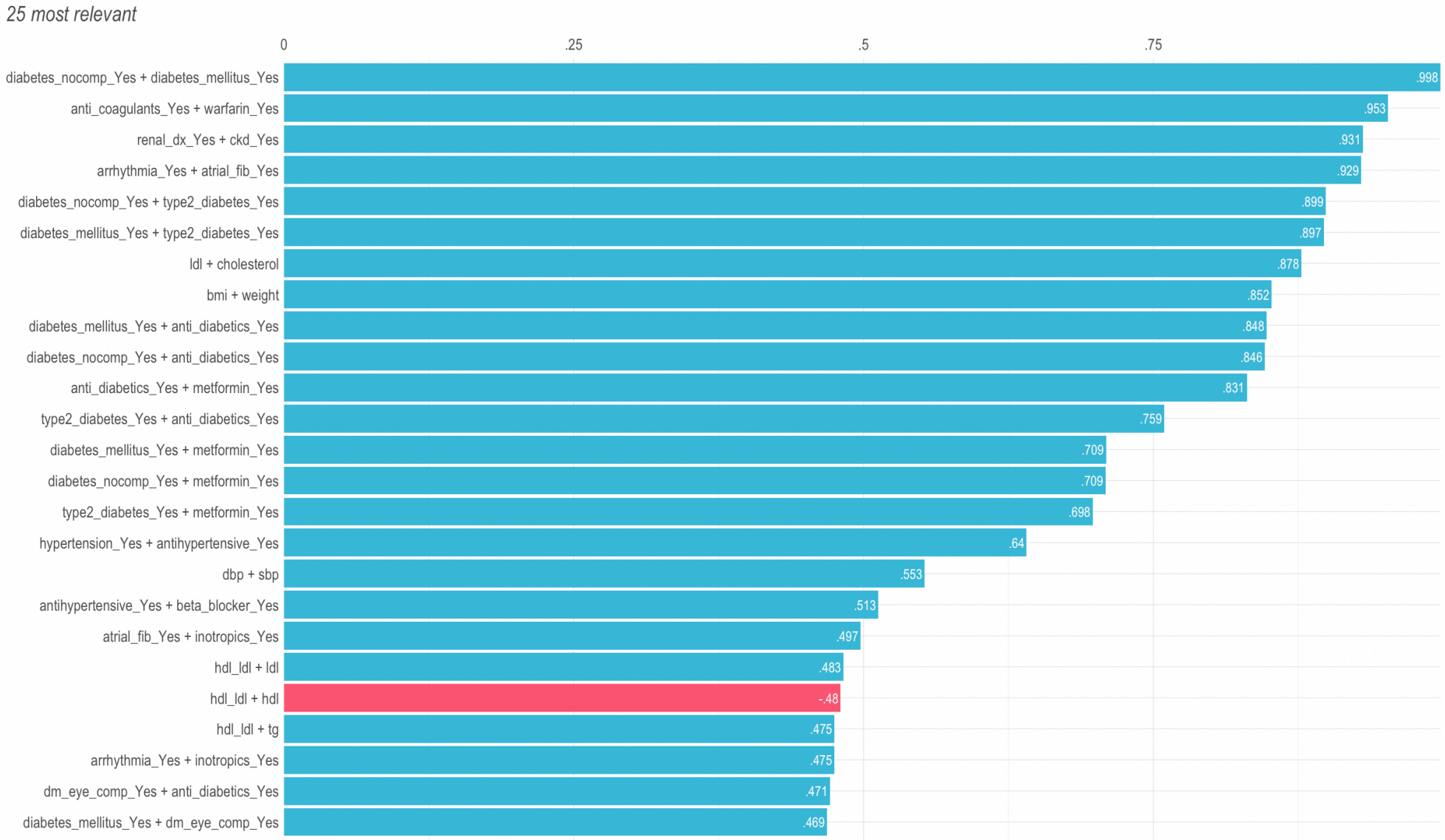


Figure 7.2 **Ranked cross-correlation plot of 49 selected variables**

Box 7.1 **Phenotypic domains and phenotypic variables used for cluster analysis**

Phenotypic domain	Phenotypic variables
Demographics	Age at incident stroke, sex, incident stroke sub-type, ethnicity, smoking status
Physical characteristics	Body mass index, diastolic and systolic blood pressures, pulse
Biochemical tests	C-reactive protein, glomerular filtration rate, haemoglobin, glycated haemoglobin, HDL cholesterol, LDL cholesterol, triglyceride
Comorbid conditions	Acute kidney injury, alcohol misuse, arrhythmia, cancer (composite), dementia, depression, diabetes mellitus (DM), DM with complications, diabetic ophthalmic complications, dyslipidaemia, hypertension, non-rheumatic aortic valve disorder, obesity, renal disease, severe mental illness, transient ischaemic attack
Prescribed medications	Anticoagulant, antidepressant, antidiabetic, antihypertensive, antiplatelet, diuretic, inotrope, loop diuretic, statin potency, thiazide diuretic

Phenotypic clustering

The prediction strength method by Tibshirani and Walther, 2015[286] in the kamila function in R and the Elbow method were used to select the optimal number of clusters – Appendix G.7 Figure 5. The kamila algorithm for mixed data clustering (Appendix G.7 Additional Methods) was implemented to identify distinct patient phenotypic clusters. To ensure the robustness of the clusters identified, 1,000 initialisations (that is, random starting points) were carried out. Plots of the clusters with the principal component analysis (PCA) dimensions was generated.

Using the h2o package in R (http://www.h2o.ai), a gradient boosting model was applied to identify as well as rank the key covariates (candidate variables) that predict each of the identified phenotypic clusters. The respective cluster groupings were coded as 1 – belonging to cluster or 0 – belonging to other clusters. SHAP

(SHapley Additive exPlanations) was used to assess the discriminative influence of the variables for each of the identified clusters.[287]

Statistical analysis

For each cluster, descriptive characteristics were provided, reporting proportion (%) for categorical variables and mean (SD) or median (IQR) for continuous variables. Kruskal-Wallis and chi-squared tests were used to compare across clusters, for continuous and categorical data, respectively.

The association between phenotypic clusters and cardiovascular morbidity and mortality outcomes were assessed using Cox proportional hazards regression model. The hazard ratio (HR) for each phenotypic group is presented with 95% confidence intervals (CI) and corresponding *p*-values. Cumulative incidence plots were derived and differences between phenotypic groups were assessed by the log-rank test. All statistical analyses were performed using Stata SE version 17 (StataCorp LP) and R version 4.1.0. An alpha level of 0.05 was used. The study flow diagram is shown in Figure 7.3.

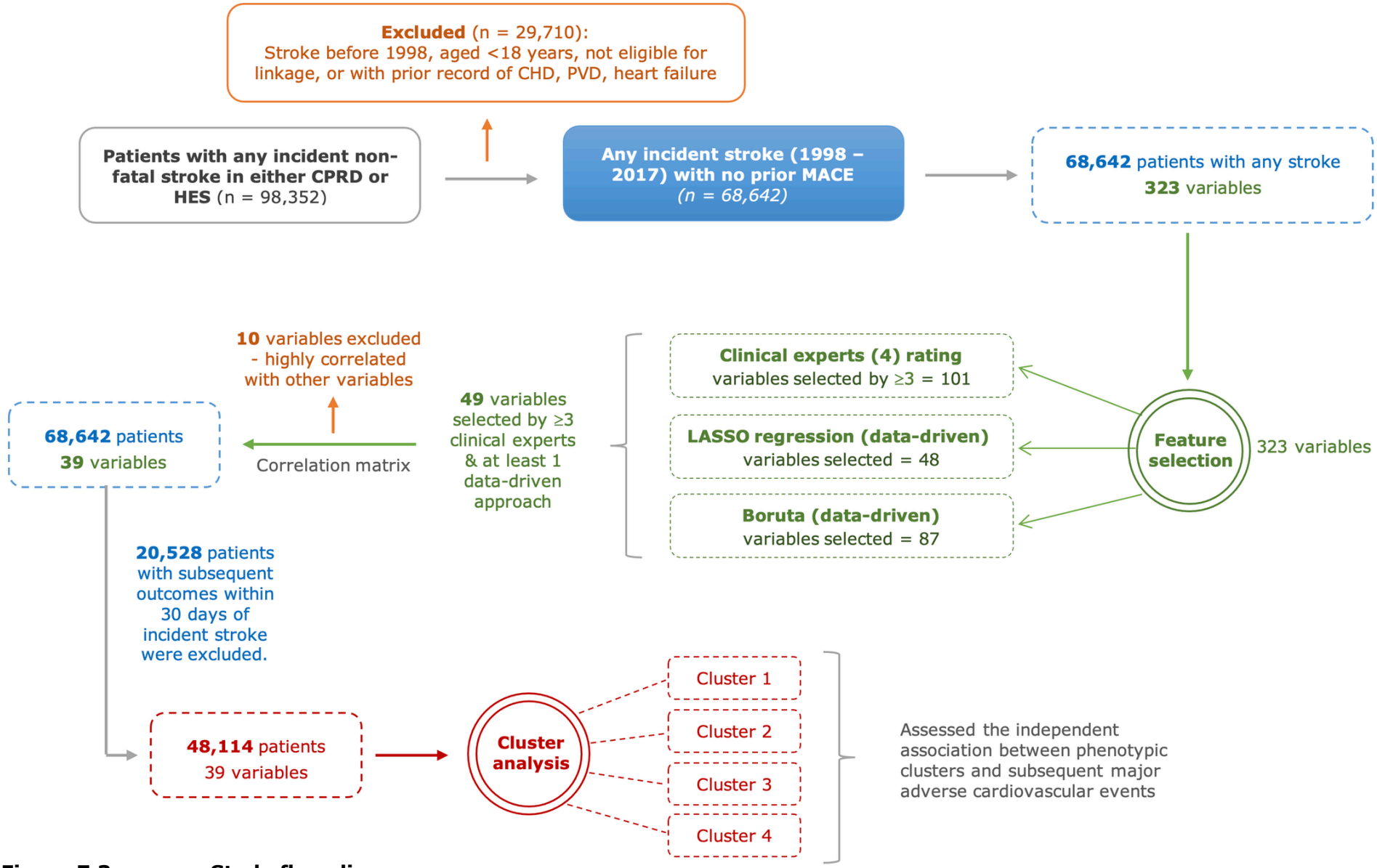


Figure 7.3 **Study flow diagram**

7.4 Results

Clinical characteristics among phenotypic clusters

There were 68,642 patients aged ≥18 years old with any incident non-fatal stroke event between 1998 and 2017 and no prior history of a serious vascular event. A total of 20,528 (29.9%) patients with subsequent clinical outcomes occurring within 30 days of the incident stroke event were excluded, as these outcomes were considered to be related to the incident stroke event.[282] Cluster analysis was performed in the remaining 48,114 patients (54.6% female). The median age for the cluster cohort was 76 years (IQR: 65–83 years). Four phenotypic clusters with significant differences in clinical characteristics were identified. The plots of the clusters are shown with the principal component analysis (PCA) dimensions in Figure 7.4. The identified clusters were numbered from 1 to 4 according to the ascendent overall incidence of the subsequent composite outcome of recurrent stroke or CVD-related mortality, the primary outcome. Table 7.2 describes and compares the clinical characteristics among the phenotypic clusters. The cluster profiles are summarised in Box 7.2.

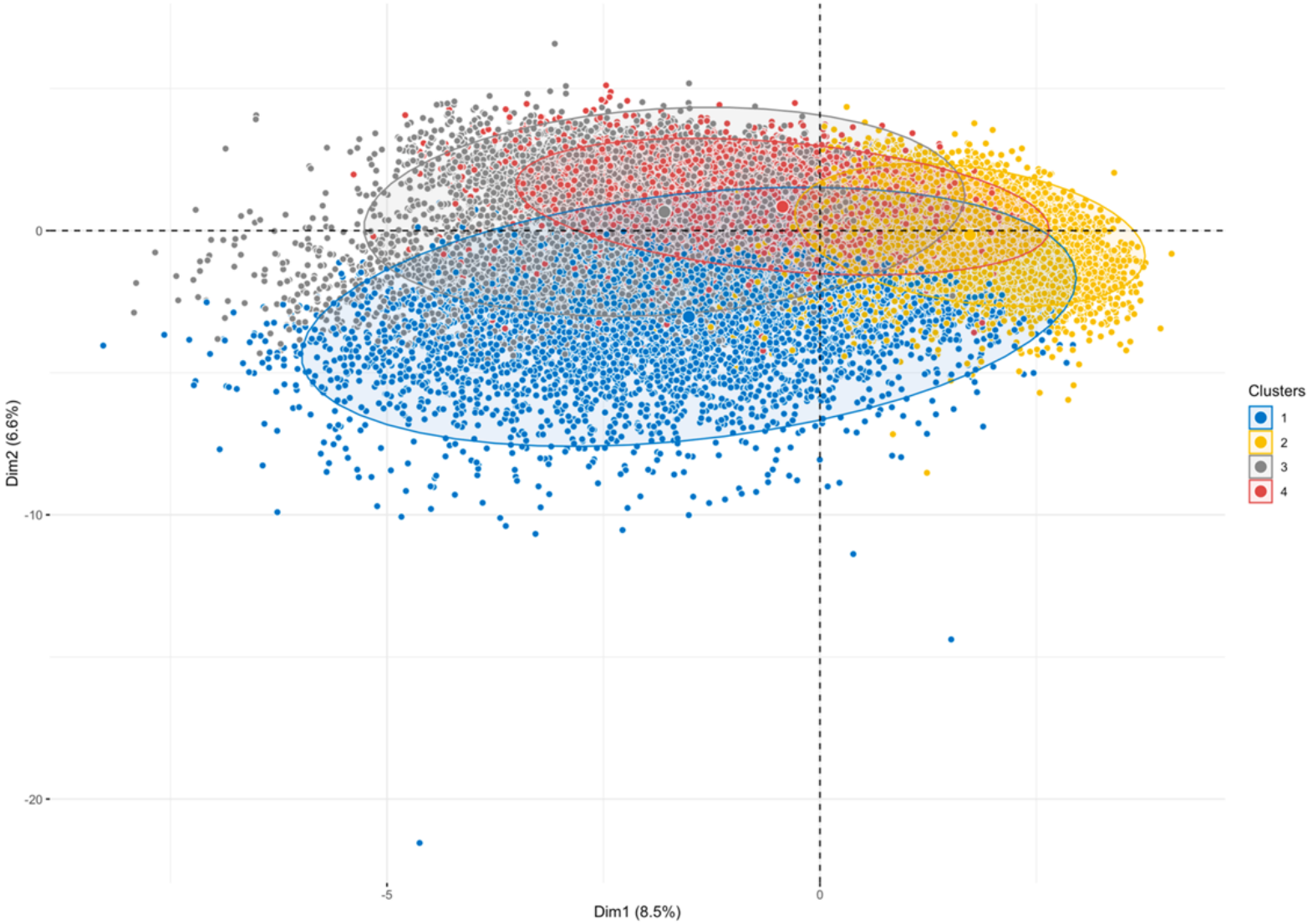


Figure 7.4 2-dimensional principal component analysis plot of clusters

Table 7.2 **Characteristics of the study population at the time of incident stroke according to cluster membership** (n=48,114)

Characteristics	**Entire cohort** 48,114 (100%)	**Cluster 1** 5,201 (10.8%)	**Cluster 2** 18,655 (38.8%)	**Cluster 3** 10,244 (21.3%)	**Cluster 4** 14,014 (29.1%)
Follow-up in years, median (IQR)	12.60 (7.60 – 16.97)	13.63 (8.67 – 17.70)	12.97 (7.97 – 17.26)	13.74 (8.81 – 17.82)	10.80 (6.02 – 15.53)
Females	26,283 (54.6)	2,120 (40.8)	8,112 (43.5)	5,490 (53.6)	10,561 (75.4)
Age (years), mean (SD)	76.0 (65.0 – 83.0)	68.0 (60.0 – 76.0)	67.0 (56.0 – 76.0)	79.0 (73.0 – 85.0)	83.0 (77.0 – 88.0)
Incident stroke subtype					
Haemorrhagic	3,336 (6.9)	216 (4.2)	1,809 (9.7)	484 (4.7)	827 (5.9)
Ischaemic	15,594 (32.4)	1,896 (36.5)	6,066 (32.5)	2,797 (27.3)	4,835 (34.5)
Stroke NOS	29,184 (60.7)	3,089 (59.4)	10,780 (57.8)	6,963 (68.0)	8,352 (59.6)
Ethnicity					
Asian	611 (1.3)	157 (3.0)	243 (1.3)	150 (1.5)	61 (0.4)
Black	377 (0.8)	87 (1.7)	140 (0.8)	69 (0.7)	81 (0.6)
Mixed	73 (0.2)	12 (0.2)	35 (0.2)	13 (0.1)	13 (0.1)
Other	335 (0.7)	50 (1.0)	152 (0.8)	66 (0.6)	67 (0.5)
White	43,011 (89.4)	4,660 (89.6)	16,589 (88.9)	9,582 (93.5)	12,180 (86.9)
Unknown	3,707 (7.7)	235 (4.5)	1,496 (8.0)	364 (3.6)	1,612 (11.5)
Socioeconomic status					
1 (Least deprived)	10,292 (21.4)	869 (16.7)	3,849 (20.6)	2,446 (23.9)	3,128 (22.3)
2	10,736 (22.3)	1,056 (20.3)	4,024 (21.6)	2,426 (23.7)	3,230 (23.0)
3	10,355 (21.5)	1,115 (21.4)	4,004 (21.5)	2,179 (21.3)	3,057 (21.8)
4	8,836 (18.4)	1,066 (20.5)	3,502 (18.8)	1,744 (17.0)	2,524 (18.0)
5 (Most deprived)	7,814 (16.2)	1,093 (21.0)	3,244 (17.4)	1,438 (14.0)	2,039 (14.5)
Unknown	81 (0.2)	2 (0.0)	32 (0.2)	11 (0.1)	36 (0.3)

Characteristics	Entire cohort 48,114 (100%)	Cluster 1 5,201 (10.8%)	Cluster 2 18,655 (38.8%)	Cluster 3 10,244 (21.3%)	Cluster 4 14,014 (29.1%)
Current smokers	8,357 (17.4)	1,247 (24.0)	4,791 (25.7)	1,054 (10.3)	1,265 (9.0)
Body mass index (kg/m^2)	26.4 (25.0 – 28.0)	30.0 (27.4 – 34.2)	26.4 (25.2 – 27.6)	25.8 (24.2 – 27.6)	26.2 (25.0 – 27.4)
DBP (mmHg)	80.0 (74.0 – 84.0)	80.0 (76.0 – 89.0)	80.0 (76.0 – 82.7)	72.0 (68.0 – 80.0)	80.0 (78.0 – 88.0)
SBP (mmHg)	140.0 (130.0 – 148.0)	142.0 (132.0 – 155.0)	139.5 (130.0 – 144.0)	133.0 (122.0 – 140.0)	145.0 (139.6 – 160.0)
C-reactive protein	9.8 (6.3 – 14.7)	9.2 (6.0 – 14.8)	10.1 (6.6 – 14.6)	8.4 (5.3 – 13.9)	10.4 (7.0 – 15.4)
Glomerular filtration rate	67.2 (62.4 – 72.0)	69.0 (61.2 – 75.0)	68.0 (64.6 – 72.5)	65.3 (58.0 – 72.0)	66.4 (61.8 – 70.4)
Glycated haemoglobin	49.9 (46.7 – 53.4)	58.3 (53.0 – 66.4)	49.7 (47.0 – 52.4)	47.5 (44.3 – 51.0)	50.2 (47.4 – 53.3)
Haemoglobin	13.5 (12.9 – 14.2)	14.2 (13.3 – 15.0)	13.6 (13.2 – 14.3)	13.2 (12.3 –14.1)	13.4 (12.7 – 13.9)
HDL cholesterol (mmol/L)	1.5 (1.3 – 1.6)	1.2 (1.0 – 1.3)	1.5 (1.3 – 1.6)	1.5 (1.3 – 1.7)	1.5 (1.4 –1.7)
LDL cholesterol (mmol/L)	3.0 (2.6 – 3.3)	3.0 (2.3 – 3.5)	3.0 (2.8 – 3.3)	2.4 (1.9 – 2.8)	3.1 (2.9 – 3.4)
Total cholesterol (mmol/L)	5.1 (4.7 – 5.4)	5.1 (4.3 – 5.8)	5.1 (4.8 – 5.4)	4.5 (3.9 – 4.9)	5.3 (5.0 – 5.7)
Triglyceride (mmol/L)	1.4 (1.2 – 1.7)	2.1 (1.6 – 2.7)	1.4 (1.3 – 1.6)	1.2 (1.0 – 1.4)	1.5 (1.3 – 1.6)
Pulse	76.4 (73.9 – 79.0)	77.8 (74.9 – 80.8)	76.6 (74.4 – 78.9)	74.8 (71.8 – 77.7)	76.7 (74.4 – 79.3)
Acute kidney injury	218 (0.5)	47 (0.9)	44 (0.2)	84 (0.8)	43 (0.3)
Alcohol problem	1,345 (2.8)	217 (4.2)	779 (4.2)	221 (2.2)	128 (0.9)
Arrhythmia	4,575 (9.5)	362 (7.0)	569 (3.1)	1,955 (19.1)	1,689 (12.1)
Atrial fibrillation	4,210 (8.8)	325 (6.3)	496 (2.7)	1,838 (17.9)	1,551 (11.1)
Cancer	7,652 (15.9)	634 (12.2)	2,167 (11.6)	2,514 (24.5)	2,337 (16.7)
Chronic kidney disease	4,945 (10.3)	767 (14.8)	390 (2.1)	2,580 (25.2)	1,208 (8.6)
Dementia	2,489 (5.2)	80 (1.5)	647 (3.5)	775 (7.6)	987 (7.0)
Depression	9,147 (19.0)	1,327 (25.5)	3,589 (19.2)	1,800 (17.6)	2,431 (17.3)
Diabetes mellitus	5,494 (11.4)	2,702 (52.0)	392 (2.1)	1,985 (19.4)	415 (3.0)
Dyslipidaemia	4,845 (10.1)	1,154 (22.2)	927 (5.0)	2,128 (20.8)	636 (4.5)
Family history of CVD	8,817 (18.3)	1,240 (23.8)	3,229 (17.3)	2,278 (22.2)	2,070 (14.8)

Characteristics	Entire cohort 48,114 (100%)	Cluster 1 5,201 (10.8%)	Cluster 2 18,655 (38.8%)	Cluster 3 10,244 (21.3%)	Cluster 4 14,014 (29.1%)
Hypertension	22,447 (46.7)	3820 (73.4)	1723 (9.2)	7885 (77.0)	9019 (64.4)
Non-rheumatic aortic valve disorder	571 (1.2)	46 (0.9)	74 (0.4)	254 (2.5)	197 (1.4)
Renal disease	5,545 (11.5)	867 (16.7)	555 (3.0)	2,764 (27.0)	1,359 (9.7)
Severe mental illness	695 (1.4)	108 (2.1)	327 (1.8)	102 (1.0)	158 (1.1)
Transient ischaemic attack	12,373 (25.7)	1,326 (25.5)	3,345 (17.9)	4,881 (47.6)	2,821 (20.1)
Anti-arrhythmic	2,163 (4.5)	227 (4.4)	451 (2.4)	698 (6.8)	787 (5.6)
Anti-coagulant	2,807 (5.8)	286 (5.5)	486 (2.6)	1,225 (12.0)	810 (5.8)
Anti-depressant	11,212 (23.3)	1,508 (29.0)	3,965 (21.3)	2,412 (23.5)	3,327 (23.7)
Anti-diabetics	4,379 (9.1)	2,476 (47.6)	254 (1.4)	1,421 (13.9)	228 (1.6)
Anti-hypertensive	23,678 (49.2)	4,231 (81.3)	2,312 (12.4)	8,497 (82.9)	8,638 (61.6)
Anti-platelet	19,789 (41.1)	2,618 (50.3)	4,605 (24.7)	6,753 (65.9)	5,813 (41.5)
Diuretics	16,835 (35.0)	2,265 (43.5)	280 (1.5)	5,288 (51.6)	9,002 (64.2)
Inotropic	2,084 (4.3)	141 (2.7)	160 (0.9)	714 (7.0)	1,069 (7.6)
Statin					
Low intensity	1,855 (3.9)	391 (7.5)	321 (1.7)	860 (8.4)	283 (2.0)
Moderate intensity	9,797 (20.4)	1,939 (37.3)	1,889 (10.1)	5,177 (50.5)	792 (5.7)
High intensity	2,240 (4.7)	713 (13.7)	335 (1.8)	1,062 (10.4)	130 (0.9)

CVD: cardiovascular disease; DBP: diastolic blood pressure; HDL: high-density lipoprotein; LDL: low-density lipoprotein; n: frequency/numbers; SBP: systolic blood pressure; SD: standard deviation; %: per cent.

Box 7.2 **Summary of cluster profiles**

Clusters	Summary description	Number (%)	Characteristic feature(s)
Cluster 1	High prevalence of CHD-related risk factors and prescribed medication	5,201 (10.8%)	Median age of 68 years (IQR 60-76), with a high proportion of patients who smoke or have diagnosed alcohol problems. Predominantly higher prevalence of CHD-related comorbidities/risk factors at the time of incident stroke – high BMI (overweight/obese), diabetes, dyslipidaemia, hypertension, and family history of CVD. A higher proportion of antidiabetic and antihypertensive prescriptions.
Cluster 2	Low prevalence of multiple long-term conditions	18,655 (38.8%)	Median age of 67 years (IQR 56-76), with lower prevalence of comorbid conditions at the time of incident stroke. A higher proportion of smokers and patients with alcohol problems. The lowest proportion of prescribed medications.
Cluster 3	High prevalence of multiple long-term conditions	10,244 (21.3%)	Median age of 79 years (IQR: 73-85) with the highest prevalence of multiple long-term conditions at the time of incident stroke – arrhythmia, cancer, chronic kidney disease, dementia, dyslipidaemia, hypertension, renal disease, and transient ischaemic attach.
Cluster 4	The oldest cohort and predominantly female	14,014 (29.1%)	The oldest cohort (median age: 83 years, IQR: 77-88) and predominantly female (75.4%). High prevalence of arrhythmia, dementia, and hypertension.

Variable importance for clusters

The supervised gradient boosting model to identify key covariates (candidate variables) that predict the respective phenotypic cluster had excellent prediction accuracy – area under the receiver operative curve (AUC[9]) of 0.985, 0.982, 0.974, and 0.970 for clusters 1, 2, 3 and 4, respectively. The most common variables for predicting the respective phenotypic clusters were age at an incident stroke, hypertension, LDL cholesterol, and potency of prescribed statin – Figure 7.5.

[9] Area under the receiver operative curve (AUC) is an overall measure of the ability to discriminate whether a specific condition or state is present or not present.[310] AUC value lies between 0.5 and 1, where 0.5 denotes a poor accuracy and 1 denotes a perfect accuracy.

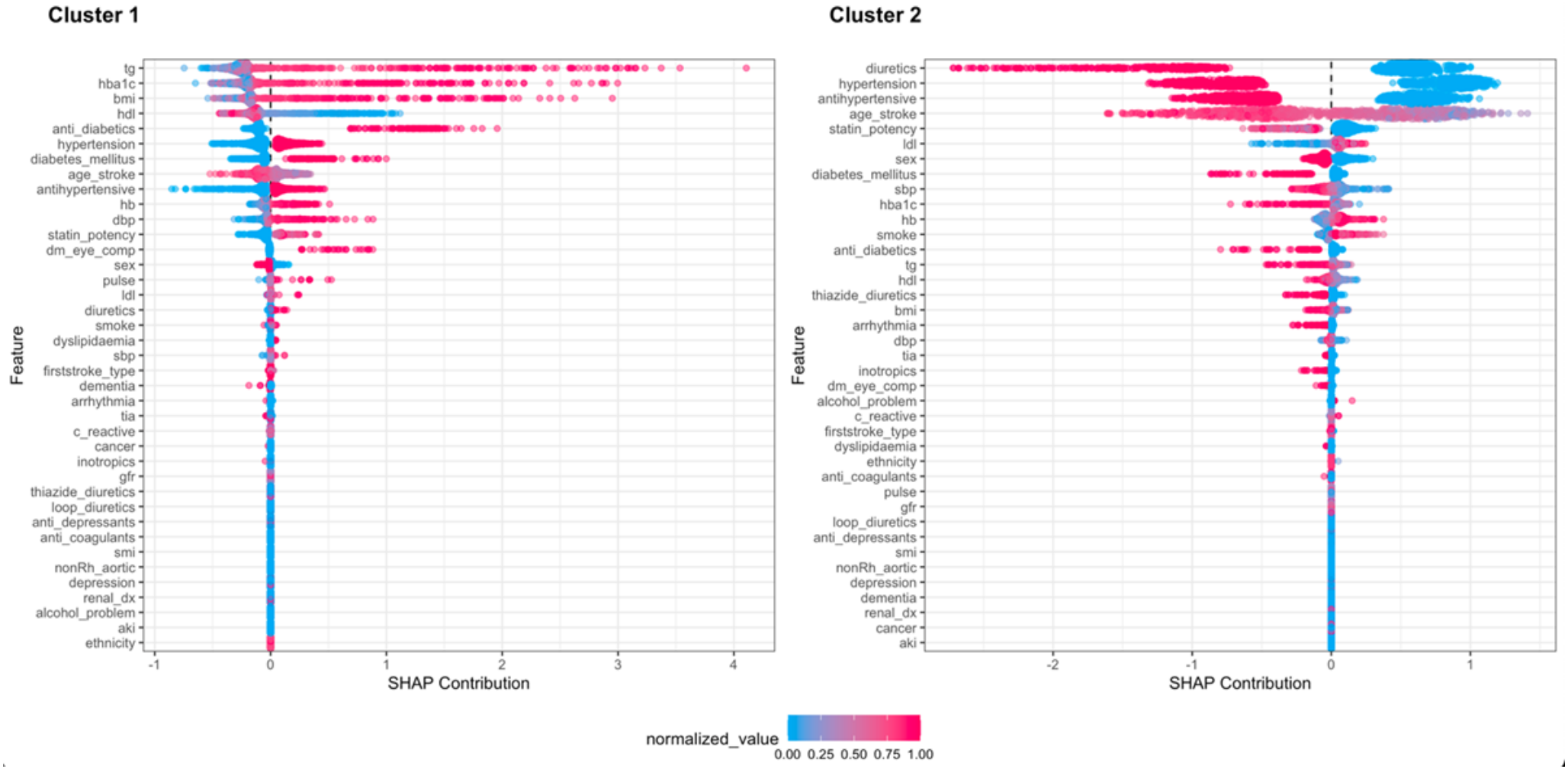


Figure 7.5 Plot showing the clinical parameters which are the core of each phenotypic cluster

aki: acute kidney injury; dbp: diastolic blood pressure; dm_eye_comp: diabetic ophthalmic complications; sbp: systolic blood pressure; gfr: glomerular filtration rate; hb: haemoglobin; hdl: high-density lipoprotein cholesterol; ldl: low-density lipoprotein cholesterol; hba1c: glycated haemoglobin; nonRH_aortic: non-rheumatic aortic valve disorder; smi: severe mental illness; tg: triglyceride; tia: transient ischaemic attack.

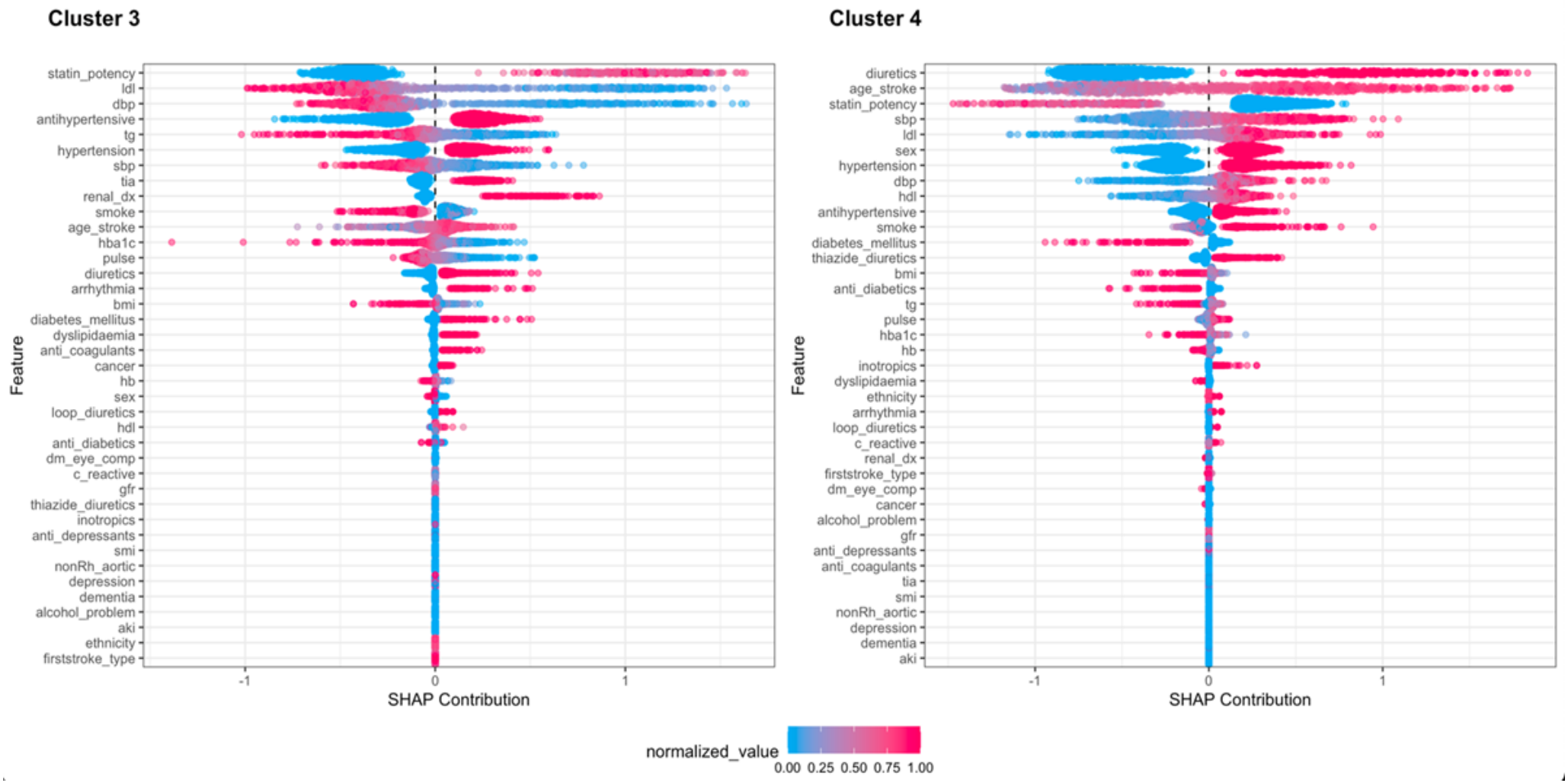


Figure 7.5 Plot showing the clinical parameters which are the core of each phenotypic cluster

SHAP summary plot combines feature/variable importance with feature effects. Each point on the summary plot is a Shapley value for an individual. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The colour represents the value from low to high. The features are ordered according to importance.

Association with subsequent clinical outcomes

During the median follow-up time of 12.60 years (IQR, 7.60 – 16.97 years), there was a total of 24,588 (51.1%) composite outcome of either recurrent stroke or CVD-related mortality events. The occurrence of recurrent stroke + CVD-related mortality was different across the 4 phenotypic clusters – cluster 1 had the lowest incidence rate (15.13 per 100 person-years; 95% CI, 14.54 – 15.74), while cluster 4 had the highest incidence rate (23.17 per 100 person-years, 95% CI: 22.67 – 23.69). The risk of subsequent recurrent stroke + CVD-related mortality was significantly increased in cluster 2 (hazard ratio (HR), 1.07; 95% CI: 1.02 – 1.12); cluster 3 (HR, 1.20; 95% CI: 1.14 – 1.26), and cluster 4 (HR, 1.29; 95% CI: 1.26 – 1.33), when compared with cluster 1. Similar incidence rate and hazard ratio trends were observed for subsequent recurrent stroke + all-cause mortality outcome (cluster 2: HR, 1.07; 95% CI, 1.03 – 1.12; cluster 3: HR, 1.32, 95% CI, 1.26 – 1.37; cluster 4: HR, 1.54; 95% CI: 1.48 – 1.60) and recurrent stroke outcome (cluster 2: HR, 1.10; 95% CI, 1.05 – 1.16; cluster 3: HR, 1.12, 95% CI, 1.06 – 1.18; cluster 4: HR, 1.25; 95% CI: 1.19 – 1.32).

Different trends in the incidence rate and hazard ratios were found/observed, however, for subsequent CHD, PVD, heart failure, CVD-related and all-cause mortality outcomes – Figure 7.6 and Table 7.3. When compared with cluster 1, the risk of subsequent CHD events was significantly decreased in the other 3 clusters (cluster 2: HR, 0.49; 95% CI: 0.44 – 0.55; cluster 3: HR, 0.64; 95% CI, 0.56 – 0.73; cluster 4: HR, 0.55; 95% CI, 0.49 – 0.63). A similar decreased risk in the other 3 clusters when compared to cluster 1 was observed for risk of subsequent PVD.

For risk of subsequent heart failure, CVD-related mortality and all-cause mortality, cluster 3 had a significantly decreased risk when compared to cluster 1 while clusters 3 and 4 had a significantly increased risk – Table 7.3. The occurrence of

subsequent cardiovascular morbidity and mortality outcomes across the different phenotypic clusters is presented as Kaplan Meier plots in Figure 7.7.

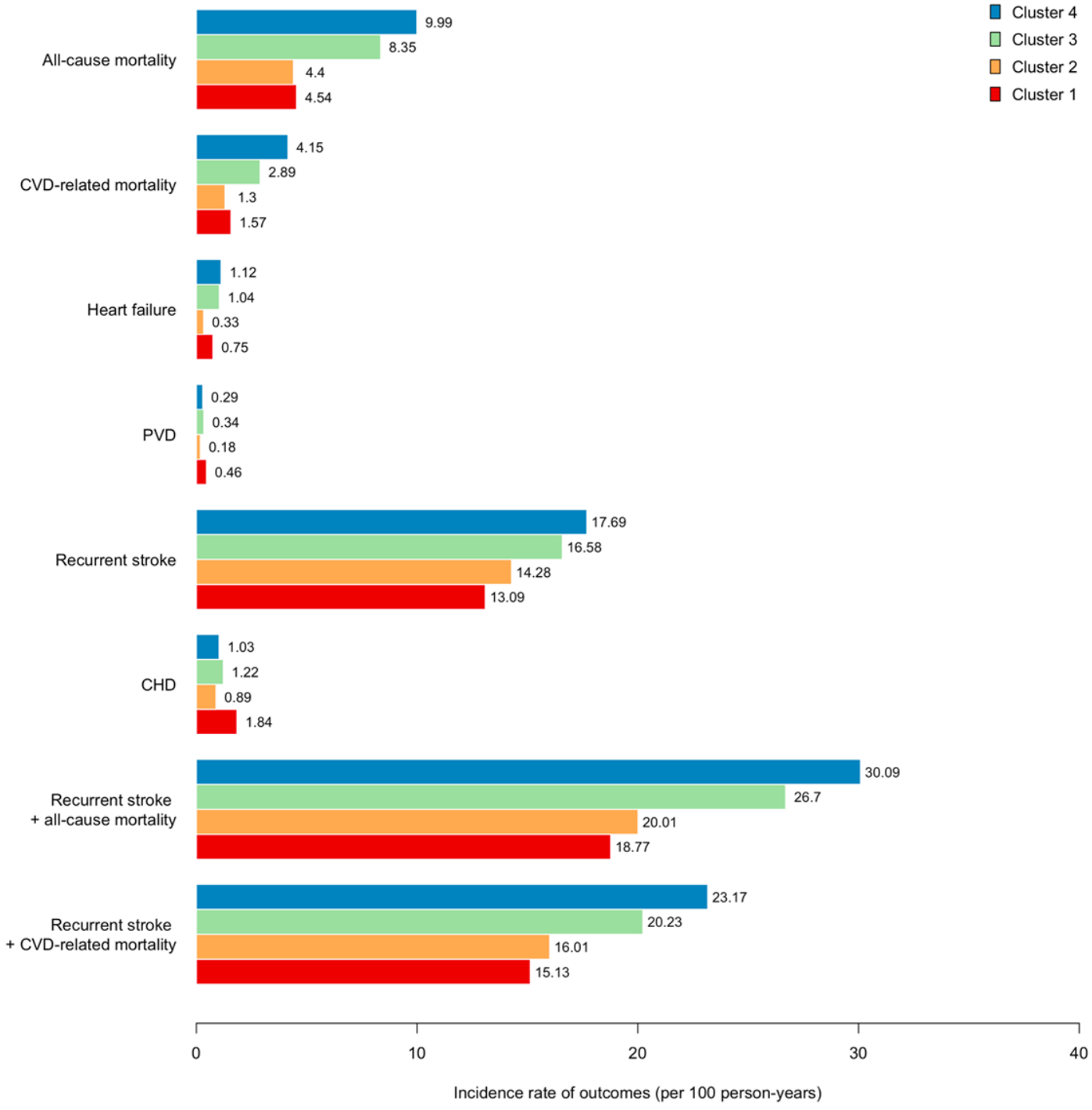


Figure 7.6 **Incidence rate for the subsequent clinical outcomes by the identified phenotypic cluster**

Table 7.3 Subsequent clinical outcomes after incident stroke by phenotypic clusters

	Number of events	Incidence rate (95% CI) per 100 PY	Hazard ratio (95% CI)
Recurrent stroke + CVD-related mortality	24,588	18.53 (18.30 – 18.76)	
Cluster 1	2,447	15.13 (14.54 – 15.74)	Reference
Cluster 2	9,249	16.01 (15.68 – 16.33)	1.07 (1.02 – 1.12)
Cluster 3	4,980	20.23 (19.68 – 20.80)	1.20 (1.14 – 1.26)
Cluster 4	7,912	23.17 (22.67 – 23.69)	1.44 (1.37 – 1.50)
Recurrent stroke + all-cause mortality	33,891	23.78 (23.52 – 24.03)	
Cluster 1	3,183	18.77 (18.13 – 19.43)	Reference
Cluster 2	12,275	20.01 (19.66 – 20.37)	1.07 (1.03 – 1.12)
Cluster 3	7,121	26.70 (26.09 – 27.33)	1.32 (1.26 – 1.37)
Cluster 4	11,312	30.09 (29.54 – 30.65)	1.54 (1.48 – 1.60)
Coronary heart disease (All)	2119	1.09 (1.04 – 1.14)	
Cluster 1	408	1.84 (1.67 – 2.02)	Reference
Cluster 2	784	0.89 (0.83 – 0.95)	0.49 (0.44 – 0.55)
Cluster 3	419	1.22 (1.10 – 1.34)	0.64 (0.56 – 0.73)
Cluster 4	508	1.03 (0.94 – 1.12)	0.55 (0.49 – 0.63)
Recurrent stroke (All)	19,810	15.42 (15.21 – 15.63)	
Cluster 1	2,075	13.09 (12.54 – 13.67)	Reference
Cluster 2	8,053	14.28 (13.97 – 14.59)	1.10 (1.05 – 1.16)
Cluster 3	3,939	16.58 (16.07 – 17.11)	1.12 (1.06 – 1.18)
Cluster 4	5,743	17.69 (17.24 – 18.15)	1.25 (1.19 – 1.32)

		Number of events	Incidence rate (95% CI) per 100 PY	Hazard ratio (95% CI)
Peripheral arterial disease (All)		529	0.27 (0.24 – 0.29)	
	Cluster 1	105	0.46 (0.38 – 0.55)	Reference
	Cluster 2	161	0.18 (0.15 – 0.21)	0.40 (0.31 – 0.51)
	Cluster 3	118	0.34 (0.28 – 0.40)	0.70 (0.54 – 0.91)
	Cluster 4	145	0.29 (0.24 – 0.34)	0.62 (0.48 – 0.79)
Heart failure (All)		1390	0.70 (0.67 – 0.74)	
	Cluster 1	172	0.75 (0.64 – 0.87)	Reference
	Cluster 2	295	0.33 (0.29 – 0.37)	0.44 (0.37 – 0.53)
	Cluster 3	363	1.04 (0.94 – 1.15)	1.34 (1.12 – 1.61)
	Cluster 4	560	1.12 (1.03 – 1.22)	1.48 (1.24 – 1.75)
Cardiovascular mortality (All)		4,778	2.34 (2.27 – 2.41)	
	Cluster 1	372	1.57 (1.42 – 1.74)	Reference
	Cluster 2	1,196	1.30 (1.22 – 1.37)	0.85 (0.75 – 0.95)
	Cluster 3	1,041	2.89 (2.72 – 3.07)	1.69 (1.50 – 1.91)
	Cluster 4	2,169	4.15 (3.98 – 4.33)	2.52 (2.25 – 2.81)
All-cause mortality (All)		14,081	6.58 (6.47 – 6.68)	
	Cluster 1	1,108	4.54 (4.28 – 4.81)	Reference
	Cluster 2	4,222	4.40 (4.27 – 4.54)	0.98 (0.92 – 1.05)
	Cluster 3	3,182	8.35 (8.06 – 8.64)	1.76 (1.64 – 1.88)
	Cluster 4	5,569	9.99 (9.73 – 10.26)	2.14 (2.01 – 2.29)

CI: confidence interval; PY: person-year

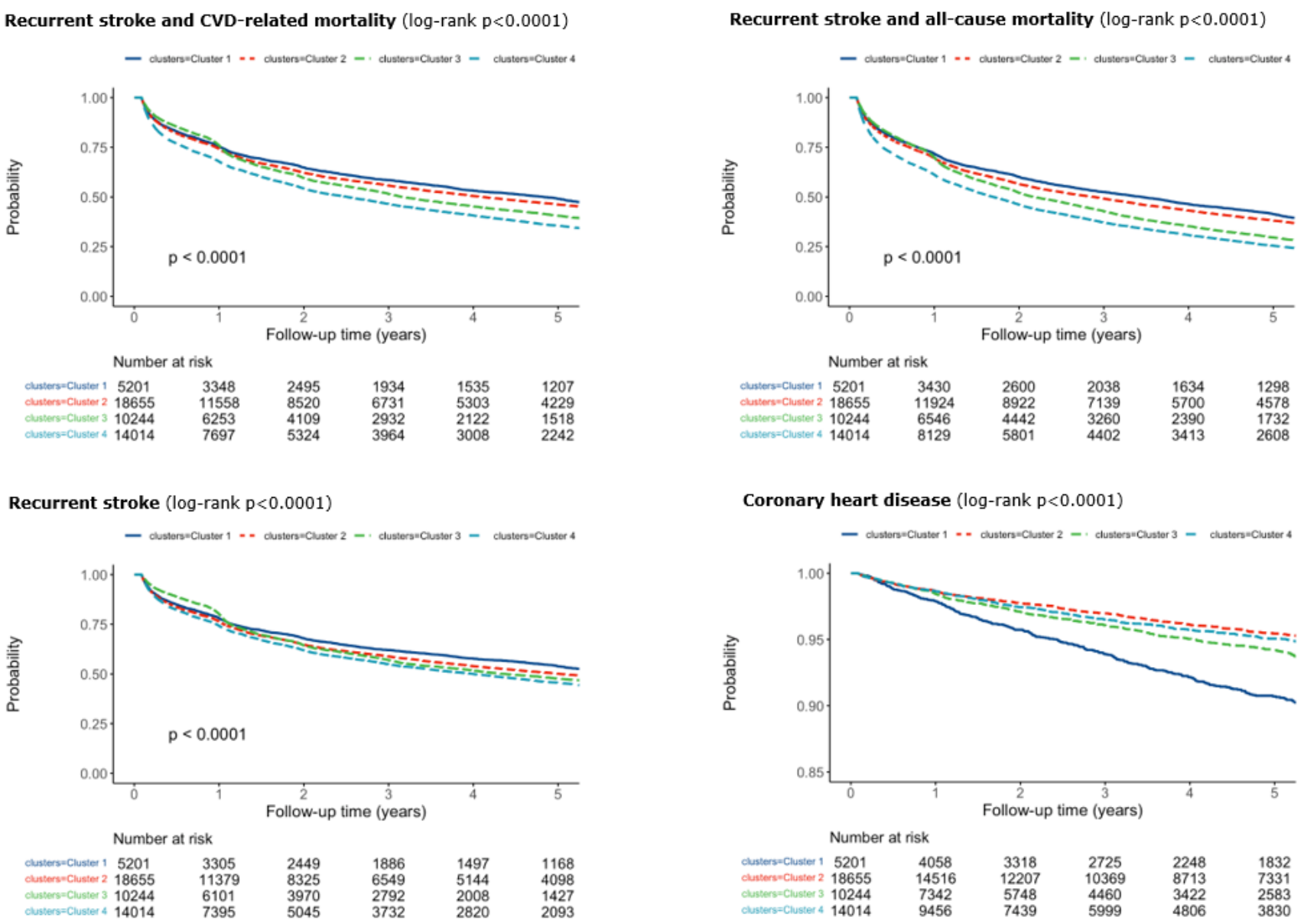


Figure 7.7 Kaplan-Meier plots for subsequent clinical outcomes stratified by phenotypic clusters

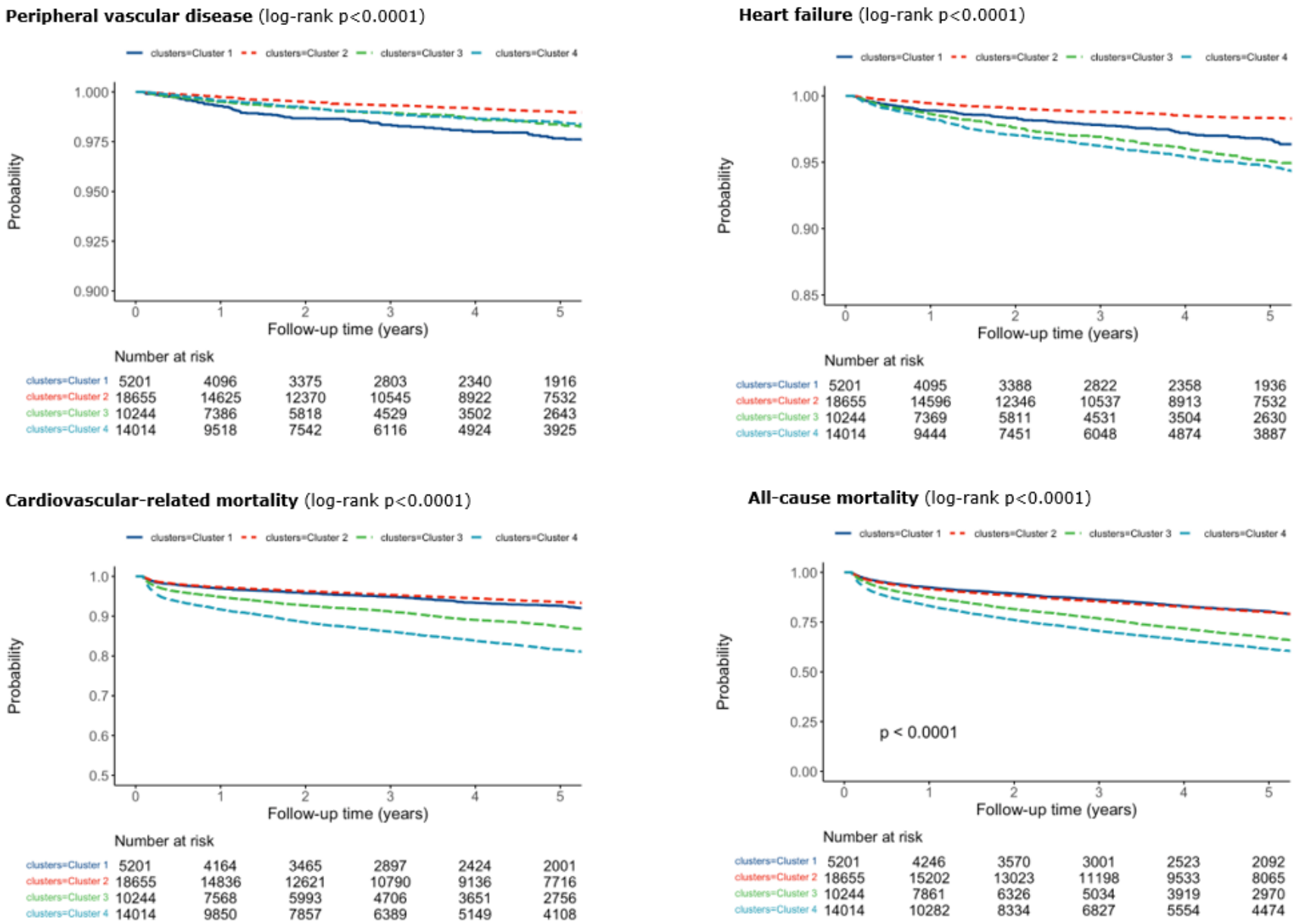


Figure 7.7 Kaplan-Meier plots for subsequent clinical outcomes stratified by phenotypic clusters

7.5 Discussion

This population-based study exploring phenotypic characteristics of patients with incident stroke using a data-driven-cluster analysis approach identified four clinically meaningful patient clusters based on the phenotypic characteristics at the time of incident stroke. Cluster 1 was a cohort with a high prevalence of CHD-related risk factors and prescribed medications; cluster 2 was a cohort with a low prevalence of multiple long-term conditions (MLTC); cluster 3 was a cohort with a high prevalence of MLTC; and cluster 4, the oldest population and predominantly female. The risk of subsequent cardiovascular morbidity and mortality outcomes differed between the identified phenotypic clusters.

In this study, four distinct and clinically meaningful phenotypic clusters were identified. Smoking, a strong independent modifiable risk factor for cardiovascular morbidity and mortality outcomes,[288] was most highly prevalent in clusters 1 and 2. A preventative strategy to communicate the risks of smoking and the benefits of quitting to this cluster of patients could be an effective means to promote smoking cessation and reduce the risk for subsequent adverse events.[289] With the exception of cluster 2, the 3 other clusters had a high prevalence of multiple long-term conditions as well as CVD risk factors at the time of incident stroke. Patients with incident stroke have been shown to commonly have pre-existing long-term conditions.[290] To optimally manage the possible atherogenic effect of these comorbid conditions to reduce the risk of subsequent cardiovascular morbidity and mortality outcomes, both non-pharmacological (that is, lifestyle modification)[76,77] and pharmacological (antihypertensives for blood pressure management;[78] lipid-lowering medications such as statins for cholesterol management;[79] antidiabetics for blood sugar control;[76] and antiplatelets/anticoagulants to manage arrhythmia[80]) strategies need to be prioritised in line with clinical guidelines.[60] Frequent monitoring/reviews to ensure treatment targets are being met is important.[81] Age, a non-modifiable risk factor, was a key factor for the patient

cluster membership. Among older adults (typical of cluster 4), the incidence of aortic disease, PVD and venous thromboembolism increase as age-related alterations in vascular structure and function are compounded by the longer exposure to CVD risk factors.[291]

Clustering is a common approach used to analyse large datasets, to identify both the number of subgroups in the data and the attributes of each subgroup, as has been done in this study. Data analysed in real applications including healthcare (from electronic health records) are mostly characterised by a mix of continuous and categorical variables (i.e., mixed-type data). More common approaches that have been applied to mixed data include converting the variables to a single data type by either coding the categorical variables as numbers or dummy coding the variables and then applying standard distance methods such as k-means designed for continuous variables to the transformed data to achieve the clustering objective(s).[292,293] Continuous variables have also been converted to categorical variables using interval-based bucketing.[294,295] Similarities that may have been observed in the original data may be lost when the data is transformed in such ways.[294] Kamila clustering algorithm has, however, been shown to better handle high imbalance between continuous and categorical data than any other method.[294,296] From a computational perspective, when compared with other algorithms, the Kamila algorithm offers the best performance and most time-efficient when dealing with large datasets (in relation to both observations and variables) in the setting of heterogeneous data, as was the situation in this study.[294,296]

Strengths and limitations

To the best of my knowledge, this is the first time a data-driven cluster analysis aimed at identifying stroke phenotypes in a well characterised large population-based cohort of adults with any incident stroke has been done. This allowed a large range of stroke phenotypes to be explored. Most importantly, I had a

comprehensive linked database with a broad spectrum of clinical data with many of these variables being explored in cluster analysis for the first time.

There are, however, limitations of this study worth considering. First and foremost, the study was not meant to propose a new classification for stroke, because the clusters are likely to vary according to patient characteristics and available data. These results serve to underscore the need for novel multidimensional stroke classification approaches for improving patient care. Furthermore, they are aimed to generate hypotheses for future studies that will integrate clinical and biological data in patients, to improve the care of patients with stroke. With immense advancement in machine learning, cluster analysis can be performed in a large number of ways.[296,297] However, the knowledge and experience of the relevant experts remain the best judge in the interpretation of findings from cluster analysis, hence the involvement of a diverse group of clinical specialists, clinical researchers, and data experts in this study.

Implications

Cluster analysis is most suited to address the multidimensional complexity of disease conditions with considerable heterogeneity such as stroke. Population-based cluster analysis could provide a further understanding of disease patterns. Additionally, patients could be phenotyped and allocated to specific clusters that could be associated with different risks for various outcomes. Different treatment strategies or interventions could be targeted at specific phenotypic clusters, based on available evidence on risk and possible response. Future clinic trial design could also focus on high-risk clusters or focus on specific aspects within a cluster.

Different types of artificial intelligence and machine learning approaches are already being used in healthcare settings – in diagnosis (e.g., aiding the work of radiologists in identifying tumours), treatment recommendation, patient engagement and adherence, and administrative activities (including guiding

researchers to construct cohorts for costly clinical trials). AI has an important role to play in the future of healthcare delivery. With respect to machine learning, it is the primary capability behind the development of precision medicine, widely agreed to be a sorely needed advance in care. Early efforts at providing diagnosis and treatment recommendations have, however, proven to be challenging. The wider adoption of AI systems will, however, need to be approved by regulators, integrated with EHR systems, standardised to a sufficient degree and updated periodically.[298]

7.6 Conclusion

Using an unsupervised machine learning cluster analysis approach, adult patients with incident stroke were grouped into four clinically meaningful phenotypic clusters based on their demographic, biochemical, comorbidities, and prescribed medication profiles at the time of incident stroke. The findings of this study highlight the significant heterogeneity that exists within patients with incident stroke with respect to subsequent cardiovascular morbidity and mortality outcomes. This offers an opportunity to revisit the stratification of care for patients with incident stroke to improve patient outcomes. The study also highlights the potential to target modifiable characteristics in clusters for more targeted preventive intervention.

Summary

This chapter used a novel cluster analysis approach for mixed-type data to classify/stratify patients with incident stroke based on sociodemographic, biochemical, comorbidity, and prescribed medication profiles, into phenotypic clusters. The differential burden of subsequent cardiovascular morbidity and mortality outcomes were evaluated.

Chapter 8

Summary conclusions and future directions for research

This concluding chapter summarises the main findings and lessons learnt from the findings, discusses the clinical as well as public health implications for the findings, and highlights opportunities for further research. The relation of my thesis studies to available evidence or wider literature has been discussed in the preceding respective study chapters.

Stroke remains a leading cause of disability and mortality globally and is associated with an economic burden. Improvements in acute care have led to many surviving after incident stroke events. As a result, a large and increasing proportion of our population is living with this long-term condition. The prognosis after surviving, however, remains sub-optimal due to the high residual risk of subsequent adverse cardiovascular morbidity and mortality outcomes. Identifying patients at the greatest risk of subsequent adverse outcomes could help clinicians and policymakers determine those patients that need to be prioritised for preventive treatment interventions.

The aim of this thesis research was, therefore, to identify clinical phenotypes (that is, patient characteristics and distinct patient clusters) that correlate with subsequent major adverse cardiovascular event outcomes (defined as a diagnosis of either CHD, recurrent stroke, PVD, heart failure, or CVD-related mortality) in adults with an incident stroke diagnosis.

In this concluding chapter, the key findings in response to the objectives are summarised in section 8.1. Section 8.2 outlines the principal strengths and limitations of this thesis research. The relation of my thesis studies to available evidence or wider literature has been discussed in the preceding respective study chapters. Section 8.3 focuses on highlighting the clinical and public health implications of the findings. Recommendations based on this thesis research for further studies are discussed in section 8.4.

8.1 Summary of main findings

Objective 1: Review of stroke prognostic models for major adverse cardiovascular outcomes

In chapter 3, the aim was to summarise the available evidence on prognostic models and evaluate their accuracy for predicting major adverse cardiovascular event outcomes in adults with an established stroke diagnosis. Forty eligible

articles with 23 distinct prognostic models were identified from 4 databases searched – describing the development of 11 prognostic models and 75 external validations. Among the 23 models, the most frequently used predictors were age, sex, history of transient ischaemic attack, hypertension (blood pressure), and diabetes. All the development models had a high risk of bias due to methodological limitations in the development of the prognostic models – improper or no information on the handling of missing data, selection of candidate predictor values, and incomplete evaluation of model performance.

Objective 2: *Describe the age, sex, and socioeconomic differences in the rates of first non-fatal stroke and subsequent major adverse outcomes*

Chapter 4 describes the age, sex, and socioeconomic differences in the incidence of non-fatal stroke events recorded between 1998 and 2017 and the incidence of subsequent major adverse outcomes. There were 82,774 incident stroke events recorded – an incidence rate of 109.20 per 100,000 person-years. The incidence of initial stroke and subsequent major adverse outcomes were higher in women, older populations, and people living in socially deprived areas.

Objective 3: *Compare the risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke*

The risk of subsequent cardiovascular morbidity and mortality outcomes between patients with incident haemorrhagic and ischaemic stroke was compared in chapter 5. Among a cohort of 32,091 patients with incident stroke (haemorrhagic stroke=6,535, ischaemic stroke=25,556), the risk of subsequent cardiovascular morbidity outcomes (CHD, recurrent stroke, PVD, and heart failure) was similar between patients with incident haemorrhagic and ischaemic stroke. Patients with

incident haemorrhagic stroke, however, had a significantly higher risk of mortality (both CVD-related and all-cause).

Objective 4: *Examine the relationship between body mass index and MACE outcomes in patients with incident stroke*

Obesity, a risk factor for stroke and is also a risk factor for hypertension and diabetes (risk factors for CVD),[82] is commonly measured using BMI. In chapter 6 I examined the relationship between BMI and subsequent cardiovascular morbidity and mortality outcomes among 30,702 patients with incident stroke and a BMI record. Patients with any type of stroke within the overweight (BMI: 25.0-29.9 kg/m^2) and obese (BMI: ≥30kg/m^2) categories were associated with a more favourable prognosis of subsequent composite MACE, PVD, and mortality (both CVD-related and all-cause) outcomes, irrespective of sex, being a smoker, or prior diagnosis of diabetes mellitus or cancer.

Objective 5: *Explore heterogeneity in clinical characteristics of adult patients with incident stroke and clustering patients based on phenotypic characteristics*

In Chapter 7 four significantly different phenotypic clusters were identified in a cohort of 48,114 patients with incident stroke and subsequent outcomes occurring 30 days after incident stroke. A novel cluster analysis approach (unsupervised machine learning technique) was used to highlight the heterogeneity of patients with incident stroke. Cluster 1 was a cohort with a high prevalence of CHD-related risk factors and prescribed medications; cluster 2 was a cohort with a low prevalence of multiple long-term conditions (MLTC); cluster 3 was a cohort with a high prevalence of MLTC; and cluster 4, the oldest population and predominantly female. The incidence and risk of subsequent cardiovascular morbidity and mortality outcomes were different for the identified phenotypic clusters.

8.2 Strengths and limitations of thesis research

Strengths and limitations have previously been highlighted in Chapter 2 (section 2.3.5) and considered in some detail in the discussion sections of the respective chapters for each of the 5 studies presented in this thesis. This concluding section, therefore, provides a summary of the principal strengths and limitations of the thesis research as a whole.

8.2.1 Strengths of the studies

The availability of anonymised and linked electronic health records across different levels of care (primary, secondary, and ambulatory care) offers a unique opportunity to better understand disease trajectories. There is the potential for improving the quality of care and outcomes for patients across different disease areas including cardiovascular disease (stroke).

The thesis studies have used a range of different data [primary care data (CPRD GOLD) to hospital (secondary care), national death records, and social deprivation data] considered reliable sources for health events; and their linkage enhances their validity and utility.[87] For instance, the completeness and accuracy of stroke recording have been shown to improve by the use of linked primary care (CPRD GOLD) and secondary care (HES APC) data.[299] The failure to use linked primary care and hospital data has been shown to lead to a substantial (25-50%) underestimation of the burden of cardiovascular disease conditions.[209]

The very large size of the linked datasets (CPRD GOLD, HES APC, ONS mortality, and social deprivation data) and the representativeness of the CPRD GOLD dataset[87] are key advantages of the datasets used. The size of the linked dataset enabled precise estimation of effect sizes and the exploration of a wide range of potential risk factors for stroke and subsequent cardiovascular morbidity and mortality outcomes. A further major strength of this research is the findings reflect

the 'real-world' context of people with incident stroke, employing their available health care data.

8.2.2 Limitations of the studies

There are some limitations associated with this thesis research worth highlighting. Firstly, the studies are reliant on the presence of clinical codes indicative of stroke, the subsequent adverse cardiovascular outcomes, and covariates (sociodemographic information, comorbid conditions, biochemical test results, and prescribed medications). Inaccurate recording may affect the estimates as is the case with all epidemiological studies using routine medical records.[300] The quality of the linked data used, however, has improved and remains high.[87]

Additionally, it is not possible to be completely certain that subsequent coding of strokes does not relate to the ongoing care of the initial stroke hence the rates for subsequent stroke may be overestimated. However, excluding patients without 12 months of data before incident stroke event sought to minimise the likelihood of overestimating stroke incidence in the thesis research studies.

Although multiple confounders were accounted for, as described, in the multivariable analyses performed there may have been other residual confounders that could have influenced the overall results of the studies. Relevant information on confounders such as lifestyle habits, diet, and physical activities levels are generally not available, that is, in being recorded sufficiently or in a consistently measurable and quantifiable way. Finally, the severity of stroke events, an important predictor of subsequent cardiovascular morbidity and mortality outcomes,[301] is not available in the linked data and hence, could not be accounted for in the analyses to estimate the risk of subsequent cardiovascular morbidity and mortality outcomes.

8.3 Clinical and public health implications

Mortality from initial strokes has seen a decline over the years, but its incidence appears to have increased.[302] The greater and growing proportion of people surviving after stroke, underlines the need to emphasize secondary prevention and enhancement of quality of life. Evidence of variation in major adverse cardiovascular event outcomes post-incident stroke by demographic and socioeconomic characteristics, therefore, offers the opportunity to tailor secondary prevention and prioritise limited healthcare resources to those at greatest risk.

The thesis research highlights women, older populations, and people in socially deprived settings as populations with a higher incidence of non-fatal strokes. Again, the same groups have a higher risk of MACE after the initial stroke episode. To address these inequalities, public health practitioners and those in primary and secondary care involved in stroke rehabilitation may utilize this evidence to design context-specific guidelines and recommendations for the management of stroke survivors, with secondary prevention measures brought to the fore. Non-pharmaceutical interventions like weight management, reduction of salt intake, smoking cessation, and implementation of healthy dietary patterns constitute the cornerstone of this holistic approach and could be targeted to the patient groups at increased risk. Focusing on intersectoral collaborations between healthcare, policy and civil groups can help implement such strategies to reduce subsequent MACE outcomes.[36]

Additionally, clinicians must focus on optimising pharmacological management of cardiovascular risk factors like arterial hypertension, high cholesterol levels, diabetes mellitus, and other strategies aimed to increase patient concordance with management and related lifestyle, as they are of paramount importance in reducing overall cardiovascular risk. Lipid-lowering treatment (including statins, ezetimibe and PCSK9 inhibitors) remains crucial for cardiovascular risk reduction,[243–246] however, statins seem to increase the risk of haemorrhagic stroke

in a dose-dependent manner, whereas PCSK9 inhibitors do not.[247] PCSK9 inhibitors may, therefore, be a preferred lipid-lowering class in patients with previous haemorrhagic stroke and particularly those at increased risk.[247] New classes of anti-thrombotic being developed offer further prospects for secondary prevention. For instance, FXIa inhibitors have shown to have a promising safety profile in preliminary reports.[242]

It is crucial for the causative mechanism of the initial stroke episode to be identified before the commencement of secondary prevention.[303] This will inform the use of surgical or non-surgical strategies or combinations of these with non-pharmacological measures. However, the multi-factorial nature of stroke episodes makes it a challenge to find default management plans to fit entire populations, hence the need for targeted interventions for different patient clusters that may experience a differing risk of adverse outcomes.

This thesis research reports findings and approaches using linked electronic health records that contribute to the evidence that may firstly inform public health policy, clinical guidelines: and secondly, inform the development and advancement of novel research methods to improve health care for the benefit of patients.

8.4 Recommendations for future studies

This thesis research presents findings that point to several areas for further research. With the Quality Outcomes Framework (QOF)[10] introduced in 2004, the quality of data recorded in UK primary care has been significantly improved and remained high especially for cardiovascular diseases and related risk factors.

[10] The pay-for-performance scheme covering a range of clinical and organisational areas in primary care

Misclassification bias[11], however, cannot be ruled out in epidemiological research using routinely collected clinical data. Validation of stroke diagnosis codes used in the primary care database remains an issue. There is, therefore, a need for more validation studies, for example, that use questionnaires sent to GPs to confirm/validate diagnosis and risk factors reported for a random sample of patients from these electronic health records.[113,114,304]

In this thesis research, a little over half of the patients identified to have an incident stroke in either primary or secondary care data had an undetermined stroke type (that is, stroke not otherwise specified). This is consistent with a recent study that assessed the completeness and accuracy of stroke recording in UK primary care.[299] The Sentinel Stroke National Audit Programme, the national stroke registry, which contains data on around 90% of all stroke hospitalisations in England and Wales, however, reported just 2% of patients with undetermined stroke type (1,076 out of 74,307 individual patient records).[305] Linkage of data to the national stroke registry as a goal, could very valuably add to the scope of information available for epidemiological research in stroke within the UK, as has been the case for myocardial infarction (MI) studies drawing additional information from the national MI register (the Myocardial Ischaemia National Audit Project [MINAP]).[88,209]

Stroke severity is an important predictor of subsequent cardiovascular morbidity and mortality outcomes in patients with a stroke event.[301] However, both primary care (CPRD GOLD) database and secondary care data (HES APC) do not have a record of the severity of stroke events recorded. Clinical guidelines recommend the use of the National Institutes of Health Stroke Scale (NIHSS) score to assess stroke severity in patients.[64] NIHSS score is a reliable predictor of outcome for

[11] Misclassification bias occurs when an individual is assigned to a different disease condition/risk factor than the one to which they should be assigned.

use in epidemiological studies.[301] The inclusion of stroke severity measure (NIHSS score) in models assessing stroke outcome is becoming a standard statistical approach in planning and implementation of stroke research.[306] The national stroke registry, which contains data on around 90% of all stroke hospitalisations in England and Wales, has the NIHSS score (stroke severity) recorded for about 73% of patients.[305] Efforts could be made to have this information made available in patients' primary care records through hospital discharge letters. This could potentially also become a QOF indicator: "*the percentage of patients with stroke diagnosis and corresponding record of stroke severity (NIHSS score) for the stroke event"*. Further studies to show the relevance of this information in providing more accurate estimates of outcomes in patients with stroke and potential cost savings as a result of having more accurate estimates and prognostic measures are needed.

Finally, in Chapter 7, my proof of principle study identified 4 distinct phenotypic clusters of adult patients with incident stroke. As indicated in Chapter 7 these findings offer an opportunity to revisit and further consider how patients are stratified based on risk for secondary prevention interventions. There is a need for these clusters to be validated and explored further in other datasets such as CPRD Aurum and Research One, to give a better understanding of stroke as a disease entity. With changes in risk factors, patient profiles, and clinical management and practice, there is a need to develop dynamic risk stratification including clustering approaches. With further development and validation, these dynamic models/approaches could be incorporated into clinical decision support tools for clinicians.

8.5 Conclusion

Huge amount of data is generated as part of health care delivery. The digitalisation and integration of routine health and care records contribute to the building of an

efficient health system that is said to promote the triple aim of better health, better health care, and lower cost. This offers new opportunities for patients to be involved in their own health and that of the general population through research. Additionally, EHRs are now important vehicles for vital research that provides insights into the quality of services as well as a better understanding of how services interact. This provides the evidence needed to inform national and international public health policies.

The studies in this thesis have demonstrated the usefulness of longitudinal linked electronic health records (UK primary care data, secondary care data, national death registry and social deprivation data) in identifying clinical phenotypes (patient characteristics and distinct patient clusters) at greater risk of subsequent major adverse cardiovascular morbidity and mortality outcomes, after the incident stroke event. Findings from this thesis research indicate patients with incident stroke experience considerable heterogeneity in clinical outcomes. The studies add to growing evidence that may inform the great potential for more accurate risk stratification of patients following an incident stroke to enable more targeted secondary prevention. Stratifying patients with stroke early, could lower the burden of subsequent adverse clinical outcomes, improve patients' long-term outcomes, and reduce the associated economic burden. This should, therefore, be a continuing research and public health priority.

References

1 Rajsic S, Gothe H, Borba HH, *et al.* Economic burden of stroke: a systematic review on post-stroke care. *Eur J Heal Econ 2018 201* 2018; **20**: 107–34.

2 King D, Wittenberg R, Patel A, Quayyum Z, Berdunov V, Knapp M. The future incidence, prevalence and costs of stroke in the UK. *Age Ageing* 2020; **49**: 277–82.

3 Wafa HA, Wolfe CDA, Emmett E, Roth GA, Johnson CO, Wang Y. Burden of Stroke in Europe. *Stroke* 2020; **51**: 2418–27.

4 NHS England. NHS Long Term Plan. NHS. 2019. https://www.longtermplan.nhs.uk/online-version/chapter-3-further-progress-on-care-quality-and-outcomes/better-care-for-major-health-conditions/cardiovascular-disease/ (accessed June 21, 2019).

5 Adams A, Bojara W, Schunk K. Early Diagnosis and Treatment of Coronary Heart Disease in Asymptomatic Subjects With Advanced Vascular Atherosclerosis of the Carotid Artery (Type III and IV b Findings Using Ultrasound) and Risk Factors. *Cardiol Res* 2018; : 22–7.

6 Ahlqvist E, Storm P, Käräjämäki A, *et al.* Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* 2018; **6**: 361–9.

7 Schiele F, Ecarnot F, Chopard R. Coronary artery disease: Risk stratification and patient selection for more aggressive secondary prevention. *Eur J Prev Cardiol* 2017; **24**: 88–100.

8 Edemekong PF, Huang B. Epidemiology Of Prevention Of Communicable Diseases. 2019 http://www.ncbi.nlm.nih.gov/pubmed/29262070 (accessed June 20, 2019).

9 World Health Organization (WHO). 10 facts on noncommunicable diseases. WHO. 2014. https://www.who.int/features/factfiles/noncommunicable_diseases/en/ (accessed June 20, 2019).

10 Engelhardt E. Apoplexy, cerebrovascular disease, and stroke: Historical evolution of terms and definitions. *Dement Neuropsychol* 2017; **11**: 449.

11 Clarke E. Apoplexy in the Hippocratic writings. *Bull Hist Med* 1963; **37**: 301–14.

12 WHO MONICA Project Principal Investigators. The World Health Organization MONICA project (monitoring trends and determinants in cardiovascular disease): A major international collaboration. *J Clin Epidemiol* 1988; **41**: 105–14.

13 Tadi P, Lui F. Acute Stroke. StatPearls. 2021. https://www.ncbi.nlm.nih.gov/books/NBK535369/ (accessed Oct 6, 2021).

14 Virani SS, Alonso A, Aparicio HJ, *et al.* Heart Disease and Stroke Statistics—2021 Update: a report from the American Heart Association. *Circulation* 2021; **143**: E254–743.

15 Chen P, Gao S, Wang Y, Xu A, Li Y, Wang D. Classifying Ischemic Stroke, from TOAST to CISS. *CNS Neurosci Ther* 2012; **18**: 456.

16 Adams H, Bendixen B, Kappelle L, *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* 1993; **24**: 35–41.

17 Campbell BC V, Khatri P. Stroke. *Lancet* 2020; **396**: 129–42.

18 Bustamante A, López-Cancio E, Pich S, *et al.* Blood Biomarkers for the Early Diagnosis of Stroke. *Stroke* 2017; **48**: 2419–25.

19 Goldstein LB, Simel DL. Is This Patient Having a Stroke? *JAMA* 2005; **293**: 2391–402.

20 Goldstein LB. Improving the Clinical Diagnosis of Stroke. *Stroke* 2006; **37**: 754–5.

21 Díez-Tejedor E, Fuentes B. Acute Care in Stroke: The Importance of Early Intervention to Achieve Better Brain Protection. *Cerebrovasc Dis* 2004; **17**: 130–7.

22 Turner AC, Zachrison KS. Utilization of Advanced Imaging for Acute Ischemic Stroke. *Circ Cardiovasc Qual Outcomes* 2021; **14**: 396–8.

23 Yew KS, Cheng E. Acute Stroke Diagnosis. *Am Fam Physician* 2009; **80**: 40.

24 Allen CM. Differential diagnosis of acute stroke: a review. *J R Soc Med* 1984; **77**: 881.

25 Katan M, Luft A. Global Burden of Stroke. *Semin Neurol* 2018; **38**: 208–11.

26 Kim J, Thayabaranathan T, Donnan GA, *et al.* Global Stroke Statistics 2019. Int. J. Stroke. 2020; **15**: 819–38.

27 Krishnamurthi R V., Ikeda T, Feigin VL. Global, Regional and Country-Specific Burden of Ischaemic Stroke, Intracerebral Haemorrhage and Subarachnoid Haemorrhage: A Systematic Analysis of the Global Burden of Disease Study 2017. *Neuroepidemiology* 2020; **54**: 171–9.

28 Sharma VK. Cerebrovascular Disease. Academic Press, 2017.

29 Mackay J, Mensah G. The Atlas of Heart Disease and Stroke. 2004. https://apps.who.int/iris/handle/10665/43007.

30 Feigin VL, Stark BA, Johnson CO, *et al.* Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol* 2021; **20**: 795–820.

31 Xu XM, Vestesson E, Paley L, *et al.* The economic burden of stroke care in England, Wales and Northern Ireland: Using a national stroke register to estimate and report patient-level health economic outcomes in stroke. *Eur Stroke J* 2018; **3**: 82–91.

32 Marmot MG, Poulter NR, Marmot MG, Poulter NR. Primary prevention of stroke. *Lancet* 1992; **339**: 344–7.

33 Boehme AK, Esenwa C, Elkind MSV. Stroke Risk Factors, Genetics, and Prevention. *Circ Res* 2017; **120**: 472–95.

34 O'Donnell MJ, Xavier D, Liu L, *et al.* Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. *Lancet* 2010; **376**: 112–23.

35 Truelsen T, Begg S, Mathers C. The global burden of cerebrovascular disease. 2006 https://www.who.int/healthinfo/statistics/bod_cerebrovasculardiseasestroke.pdf (accessed Sept 20, 2021).

36 Pandian JD, Gall SL, Kate MP, *et al.* Prevention of stroke: a global perspective. *Lancet* 2018; **392**: 1269–78.

37 Go AS, Mozaffarian D, Roger VL, *et al.* Executive summary: Heart Disease and Stroke Statistics - 2014 Update: A report from the American Heart Association. Circulation. 2014; **129**: 399–410.

38 Johnson CO, Nguyen M, Roth GA, *et al.* Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* 2019; **18**: 439–58.

39 Cipolla MJ, Liebeskind DS, Chan S-L. The importance of comorbidities in ischemic stroke: Impact of hypertension on the cerebral circulation. *J Cereb Blood Flow Metab* 2018; **38**: 2149.

40 Checchin D, Freeman MA, Terres JAR. Comorbid Conditions of Patients Hospitalized for Stroke in Canada, 2009/2010. *Can J Cardiol* 2012; **28**: S220–1.

41 Magwood GS, White BM, Ellis C. Stroke-Related Disease Comorbidity and Secondary Stroke Prevention Practices among Young Stroke Survivors. *J Neurosci Nurs* 2017; **49**: 296–301.

42 British Heart Foundation. Incidence and prevalence - comorbidities (stroke). 2018. https://www.bhf.org.uk/what-we-do/our-research/heart-and-circulatory-diseases-in-numbers/comorbidities-stroke (accessed Sept 19, 2021).

43 Kuriakose D, Xiao Z. Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives. *Int J Mol Sci* 2020; **21**: 7609.

44 Gallacher KI, McQueenie R, Nicholl B, Jani BD, Lee D, Mair FS. Risk Factors and Mortality Associated with Multimorbidity in People with Stroke or Transient Ischaemic Attack: A Study of 8,751 UK Biobank Participants. *J Comorbidity* 2018; **8**: 1–8.

45 Intercollegiate Stroke Working Party. National Clinical Guideline for Stroke. 2016.

46 Stroke Unit Trialists' Collaboration. Organised inpatient (stroke unit) care for stroke. *Cochrane Database Syst Rev* 2013; **2013**. DOI:10.1002/14651858.CD000197.PUB3.

47 Rodgers H, Price C. Stroke unit care, inpatient rehabilitation and early supported discharge. *Clin Med (Northfield Il)* 2017; **17**: 173.

48 NHS Wales Informatics Service. Health in Wales: Acute Stroke. 2010. http://www.wales.nhs.uk/acutestroke (accessed Oct 6, 2021).

49 Leavell HR, Clark EG. Preventive medicine for the doctor in his community : an epidemiologic approach. New York: McGraw-Hill, 1965.

50 Goldston SE. Concepts of primary prevention: a framework for program development. Sacramento: Department of Mental Health, Office of Prevention, 1987.

51 Australian Institute of Health and Welfare. Prevention of cardiovascular disease, diabetes and chronic kidney disease: targeting risk factors. Canberra: AIHW., 2009 www.aihw.gov.au (accessed June 21, 2019).

52 Wallace RB. Tertiary Prevention. Encyclopedia.com. 2007. https://www.encyclopedia.com/medicine/divisions-diagnostics-and-procedures/medicine/tertiary-prevention (accessed June 21, 2019).

53 Thrombosis Adviser. Resource about Venous & Arterial Thrombosis. 2019. https://www.thrombosisadviser.com/coronary-and-peripheral-artery-disease/ (accessed June 21, 2019).

54 Silver B. Stroke Prevention. Medscape. 2021. https://emedicine.medscape.com/article/323662-overview#a4 (accessed Sept 19, 2021).

55 World Health Organization. Prevention of recurrent heart attacks and strokes in low and middle income populations: evidence-based recommendations for policy-makers and health professionals. 2004. https://apps.who.int/iris/handle/10665/42842.

56 Albert MA, Glynn RJ, Buring J, Ridker PM. Impact of Traditional and Novel Risk Factors on the Relationship Between Socioeconomic Status and Incident Cardiovascular Events. *Circulation* 2006; **114**: 2619–26.

57 National Institute for Health and Care Excellence. Cardiovascular Disease Prevention. London: National Institute for Health and Care Excellence, 2010 https://www.nice.org.uk/guidance/ph25 (accessed Aug 3, 2019).

58 Challa HJ, Ameer MA, Uppaluri KR. DASH Diet To Stop Hypertension. Treasure Island (FL): StatPearls Publishing, 2021 https://www.ncbi.nlm.nih.gov/books/NBK482514/ (accessed Sept 19, 2021).

59 National Institute for Health and Care Excellence. Scenario: Secondary prevention following stroke and TIA . 2021. https://cks.nice.org.uk/topics/stroke-tia/management/secondary-prevention-following-stroke-tia/ (accessed Jan 5, 2022).

60 Kleindorfer DO, Towfighi A, Chaturvedi S, *et al.* 2021 Guideline for the Prevention of Stroke in Patients With Stroke and Transient Ischemic Attack: A Guideline From the American Heart Association/American Stroke Association. *Stroke* 2021; **52**: E364–467.

61 Piepoli MF, Hoes AW, Agewall S, *et al.* 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice. *Eur Heart J* 2016; **37**: 2315–81.

62 Smith SC, Benjamin EJ, Bonow RO, *et al.* AHA/ACCF Secondary Prevention and Risk Reduction Therapy for Patients With Coronary and Other Atherosclerotic Vascular Disease: 2011 Update. *Circulation* 2011; **124**: 2458–73.

63 Grundy SM, Stone NJ. 2018 Cholesterol Clinical Practice Guidelines: Synopsis of the 2018 American Heart Association/American College of Cardiology/Multisociety Cholesterol Guideline. *Ann Intern Med* 2019; **170**: 779.

64 National Institute for Health and Care Excellence. Stroke and transient ischaemic attack in over 16s: diagnosis and initial management. 2019 www.nice.org.uk/guidance/ng128 (accessed Jan 17, 2022).

65 LaRosa JC, Grundy SM, Kastelein JJP, Kostis JB, Greten H. Safety and Efficacy of Atorvastatin-Induced Very Low-Density Lipoprotein Cholesterol Levels in Patients With Coronary Heart Disease (a Post Hoc Analysis of the Treating to New Targets [TNT] Study). *Am J Cardiol* 2007; **100**: 747–52.

66 The Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) Investigators. High-Dose Atorvastatin after Stroke or Transient Ischemic Attack. *N Engl J Med* 2006; **355**: 549–59.

67 Catapano AL, Graham I, De Backer G, *et al.* 2016 ESC/EAS Guidelines for the Management of Dyslipidaemias. *Eur Heart J* 2016; **37**: 2999–3058.

68 Law MR, Morris JK, Wald NJ. Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. *BMJ* 2009; **338**: b1665.

69 Antithrombotic Trialists' (ATT) Collaboration AT (ATT), Baigent C, Blackwell L, *et al.* Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. *Lancet* 2009; **373**: 1849–60.

70 Baigent C, Blackwell L, Emberson J, *et al.* Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170 000 participants in 26 randomised trials. *Lancet* 2010; **376**: 1670–81.

71 Kotseva K, Wood D, Backer G De, Bacquer D De, Pyörälä K, Keil U. EUROASPIRE III: a survey on the lifestyle, risk factors and use of cardioprotective drug therapies in coronary patients from 22 European countries. *Eur J Cardiovasc Prev Rehabil* 2009; **16**: 121–37.

72 Eesa M, Schumacher HC, Higashida RT, Meyers PM. Advances in revascularization for acute ischemic stroke treatment: An update. Expert Rev. Neurother. 2011; **11**: 1125–39.

73 Pierot L, Soize S, Benaissa A, Wakhloo AK. Techniques for Endovascular Treatment of Acute Ischemic Stroke. *Stroke* 2015; **46**: 909–14.

74 Barnes DE, Yaffe K. The projected effect of risk factor reduction on Alzheimer's disease prevalence. *Lancet Neurol* 2011; **10**: 819–28.

75 Kelly DM, Rothwell PM. Impact of multimorbidity on risk and outcome of stroke: Lessons from chronic kidney disease. *Int J Stroke* 2021; **16**: 758–70.

76 Kernan WN, Ovbiagele B, Black HR, *et al.* Guidelines for the Prevention of Stroke in Patients With Stroke and Transient Ischemic Attack. *Stroke* 2014; **45**: 2160–236.

77 Billinger SA, Arena R, Bernhardt J, *et al.* Physical activity and exercise recommendations for stroke survivors: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2014; **45**: 2532–53.

78 Arima H, Chalmers J, Woodward M, *et al.* Lower target blood pressures are safe and effective for the prevention of recurrent stroke: The PROGRESS trial. *J Hypertens* 2006; **24**: 1201–8.

79 Fulcher J, O'Connell R, Voysey M, *et al.* Efficacy and safety of LDL-lowering therapy among men and women: meta-analysis of individual data from 174 000 participants in 27 randomised trials. *Lancet* 2015; **385**: 1397–405.

80 CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *Lancet* 1996; **348**: 1329–39.

81 National Institute for Health and Care Excellence. Multimorbidity: clinical assessment and management (NG56). 2016. https://www.nice.org.uk/guidance/ng56 (accessed Oct 1, 2021).

82 Suk SH, Sacco RL, Boden-Albala B, *et al.* Abdominal obesity and risk of ischemic stroke: the Northern Manhattan Stroke Study. *Stroke* 2003; **34**: 1586–92.

83 Moher D, Shamseer L, Clarke M, *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; **4**: 1.

84 Vandenbroucke JP, Elm E von, Altman DG, *et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLOS Med* 2007; **4**: e297.

85 Health and Social Care Information Centre. Attribution Data Set GP-Registered Populations Scaled to ONS Population Estimates - 2011. 2012. https://digital.nhs.uk/data-and-information/publications/statistical/attribution-dataset-gp-registered-populations/attribution-data-set-gp-registered-populations-scaled-to-ons-population-estimates-2011 (accessed Sept 6, 2021).

86 Wolf A, Dedman D, Campbell J, *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019. DOI:10.1093/ije/dyz034.

87 Herrett E, Gallagher AM, Bhaskaran K, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44**: 827–36.

88 Denaxas SC, George J, Herrett E, *et al.* Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 2012; **41**: 1625–38.

89 Lester H, Campbell S. Developing Quality and Outcomes Framework (QOF) indicators and the concept of 'QOFability'. *Qual Prim Care* 2010; **18**: 103–9.

90 Mathur R, Bhaskaran K, Chaturvedi N, *et al.* Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health (Bangkok)* 2014; **36**: 684–92.

91 Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017; **46**: 1093-1093i.

92 National Audit Office. Healthcare across the UK: A comparison of the NHS in England, Scotland, Wales and Northern Ireland. 2012 https://www.nao.org.uk/wp-content/uploads/2012/06/1213192.pdf (accessed Aug 31, 2021).

93 CDC/National Center for Health Statistics. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). Centers Dis. Control Prev. 2019. https://www.cdc.gov/nchs/icd/icd10cm.htm (accessed June 21, 2019).

94 NHS Digital. OPCS Classification of Interventions and Procedures. 2021. https://datadictionary.nhs.uk/supporting_information/opcs_classification_of_interventions_and_procedures.html (accessed Aug 31, 2021).

95 Office for National Statistics. Deaths Registration Data. ONS. 2018. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths (accessed June 21, 2019).

96 Health & Social Care Information Centre. A Guide to Linked Mortality Data from Hospital Episode Statistics and the Office for National Statistics. 2015 https://webarchive.nationalarchives.gov.uk/ukgwa/20180307213524tf_/http://content.digital.nhs.uk/media/11668/HES-ONS-Mortality-Data-Guide/pdf/mortality_guide.pdf (accessed Aug 31, 2021).

97 Department of Communities and Local Government. English Indices of Deprivation 2015. 2015; **1**: 1–11.

98 Noble M, Wright G, Smith G, Dibben C. Measuring Multiple Deprivation at the Small-Area Level. *Environ Plan A Econ Sp* 2006; **38**: 169–85.

99 Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2005; **14**: 443–51.

100 Kuan V, Denaxas S, Gonzalez-Izquierdo A, *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Heal* 2019; **1**: e63–77.

101 CPRD @ Cambridge. Codes Lists (GOLD). https://www.phpc.cam.ac.uk/pcu/research/research-groups/crmh/cprd_cam/codelists/v11/ (accessed March 6, 2021).

102 Barnett V, Lewis T. Outliers in statistical data. Chichester: Wiley, 1994.

103 Osborne J, Overbay A. The power of outliers (and why researchers should always check for them). *Pract Assessment, Res Eval* 2004; **9**: 6.

104 Zhao Y. Chapter 7 - Outlier Detection. In: Zhao YBT-R and DM, ed. R and Data Mining. Academic Press, 2013: 63–73.

105 Ghosh D, Vogt A. Outliers: An Evaluation of Methodologies. In: Survey Research Methods - Joint Statistical Meetings. 2012: 3455–60.

106 Little RJA, Rubin DB. Statistical Analysis with Missing Data. Wiley, 2002 DOI:10.1002/9781119013563.

107 Sterne JAC, White IR, Carlin JB, *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**: b2393.

108 Royston P. Multiple imputation of missing values: Update of ice. *Stata J* 2005; **5**: 527–36.

109 Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001; **6**: 330–51.

110 Kontopantelis E, White IR, Sperrin M, Buchan I. Outcome-sensitive multiple imputation: a simulation study. *BMC Med Res Methodol* 2017; **17**: 1–13.

111 Rubin DB. Multiple imputation for nonresponse in surveys. Wiley, 1987 DOI:10.1002/9780470316696.

112 Strongman H, Williams R, Meeraus W, *et al.* Limitations for health research with restricted data collection from UK primary care. *Pharmacoepidemiol Drug Saf* 2019; **28**: 777–87.

113 Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: A systematic review. *Br J Clin Pharmacol* 2010; **69**: 4–14.

114 Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: A systematic review. Br. J. Gen. Pract. 2010; **60**: 199–206.

115 Randhawa G. Tackling health inequalities for minority ethnic groups: challenges and opportunities. Briefing 6. 2007 https://raceequalityfoundation.org.uk/wp-content/uploads/2018/03/health-brief6.pdf (accessed May 29, 2022).

116 Routen A, O'Mahoney L, Ayoubkhani D, *et al.* Understanding and tracking the impact of long COVID in the United Kingdom. *Nat Med* 2021; **28**: 11–5.

117 Gopalakrishnan S, Ganeshkumar P. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *J Fam Med Prim Care* 2013; **2**: 9.

118 Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997; **126**: 376–80.

119 Biermans MCJ, Verheij RA, Bakker DH de, Zielhuis GA, Robbé PF de V. Estimating Morbidity Rates from Electronic Medical Records in General Practice. *Methods Inf Med* 2018; **47**: 98–106.

120 Williams R, Wright J. Health needs assessment: Epidemiological issues in health needs assessment. *Br Med J* 1998; **316**: 1382.

121 Giampaoli S, Palmieri L, Capocaccia R, Pilotto L, Vanuzzo D. Estimating population-based incidence and prevalence of major coronary events. *Int J Epidemiol* 2001; **30**: S5-10.

122 Keiding N. Age-Specific Incidence and Prevalence: A Statistical Perspective. *J R Stat Soc Ser A (Statistics Soc* 1991; **154**: 371–96.

123 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**: 37–48.

124 Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011; **46**: 424.

125 Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983; **70**: 41–55.

126 Rosenbaum PR, Rubin DB. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *J R Stat Soc Ser B* 1983; **45**: 212–8.

127 Morgan CJ. Landmark analysis: A primer. *J Nucl Cardiol* 2019; **26**: 391–3.

128 Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol* 1983; **1**: 710–9.

129 Jones M, Fowler R. Immortal time bias in observational studies of time-to-event outcomes. *J Crit Care* 2016; **36**: 195–9.

130 Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B* 1972; **34**: 187–202.

131 Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *Br J Cancer* 2003; **89**: 431.

132 Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. *Br J Cancer* 2003; **89**: 232–8.

133 Grant SW, Hickey GL, Head SJ. Statistical primer: multivariable regression considerations and pitfalls. *Eur J Cardio-Thoracic Surg* 2019; **55**: 179–85.

134 Frades I, Matthiesen R. Overview on Techniques in Cluster Analysis. In: Methods in Molecular Biology. Humana Press, 2010: 81–107.

135 Eick CF, Zeidat N, Zhao Z. Supervised clustering - Algorithms and benefits. In: Proceedings - International Conference on Tools with Artificial Intelligence (ICTAI). 2004: 774–6.

136 Bandeira LPC, Sousa JMC, Kaymak U. Fuzzy Clustering in Classification Using Weighted Features. In: Fuzzy Sets and Systems — IFSA 2003. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). Springer, Berlin, Heidelberg, 2003: 560–7.

137 Ester M, Kriegel HP, Sander J, Xiaowei X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. United States: AAAI Press, Menlo Park, CA (United States), 1996: 226–231.

138 Rehioui H, Idrissi A, Abourezq M, Zegrari F. DENCLUE-IM: A New Approach for Big Data Clustering. *Procedia Comput Sci* 2016; **83**: 560–7.

139 Rendón E, Abundez I, Arizmendi A, Quiroz EM. Internal versus External cluster validation indexes. *Int J Comput Commun* 2011; **5**.

140 Roth GA, Johnson C, Abajobir A, *et al.* Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol* 2017; **70**: 1–25.

141 Mozaffarian D. Global Scourge of Cardiovascular Disease: Time for Health Care Systems Reform and Precision Population Health. *J Am Coll Cardiol* 2017; **70**: 26–8.

142 Rosamond WD, Chambless LE, Folsom AR, *et al.* Trends in the Incidence of Myocardial Infarction and in Mortality Due to Coronary Heart Disease, 1987 to 1994. *N Engl J Med* 1998; **339**: 861–7.

143 Chen H-Y, Gore JM, Lapane KL, *et al.* A 35-Year Perspective (1975 to 2009) into the Long-Term Prognosis and Hospital Management of Patients Discharged from the Hospital After a First Acute Myocardial Infarction. *Am J Cardiol* 2015; **116**: 24–9.

144 Mohan KM, Wolfe CDA, Rudd AG, Heuschmann PU, Kolominsky-Rabas PL, Grieve AP. Risk and cumulative risk of stroke recurrence: A systematic review and meta-analysis. *Stroke* 2011; **42**: 1489–94.

145 Johnston SC, Gress DR, Browner WS, Sidney S. Short-term prognosis after emergency department diagnosis of TIA. *J Am Med Assoc* 2000; **284**: 2901–6.

146 Riley RD, Hayden JA, Steyerberg EW, *et al.* Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Med* 2013; **10**: e1001380.

147 Moons KGM, de Groot JAH, Bouwmeester W, *et al.* Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Med* 2014; **11**: e1001744.

148 Page MJ, McKenzie JE, Bossuyt PM, *et al.* The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 2021; **372**. DOI:10.1136/bmj.n71.

149 Akyea R, Leonardi-Bee J, Asselbergs F, *et al.* Predicting major adverse cardiovascular events for secondary prevention: protocol for a systematic review and meta-analysis of risk prediction models. *BMJ Open* 2020; **10**: e034564.

150 Moons KGM, Kengne AP, Grobbee DE, *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012; **98**: 691–8.

151 Wolff RF, Moons KGM, Riley RD, *et al.* PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019; **170**: 51.

152 Dorresteijn JAN, Visseren FLJ, Wassink AMJ, *et al.* Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. *Heart* 2013; **99**: 866.

153 Andersen SD, Gorst-Rasmussen A, Lip GY, Bach FW, Larsen TB. Recurrent Stroke: The Value of the CHA2DS2VASc Score and the Essen Stroke Risk Score in a Nationwide Stroke Cohort. *Stroke* 2015; **46**: 2491–7.

154 Andersen SD, Larsen TB, Gorst-Rasmussen A, Yavarian Y, Lip GYH, Bach FW. White Matter Hyperintensities Improve Ischemic Stroke Recurrence Prediction. *Cerebrovasc Dis* 2017; **43**: 17–24.

155 Arsava EM, Furie KL, Schwamm LH, Sorensen AG, Ay H. Prediction of early stroke risk in transient symptoms with infarction: Relevance to the new tissue-based definition. *Stroke* 2011; **42**: 2186–90.

156 Arsava EM, Kim G-M, Oliveira-Filho J, *et al.* Prediction of Early Recurrence After Acute Ischemic Stroke. *JAMA Neurol* 2016; **73**: 396–401.

157 Asimos AW, Johnson AM, Rosamond WD, *et al.* A multicenter evaluation of the ABCD2 score's accuracy for predicting early ischemic stroke in admitted patients with transient ischemic attack. *Ann Emerg Med* 2010; **55**: 201–10.

158 Ay H, Gungor L, Arsava EM, *et al.* A score to predict early risk of recurrence after ischemic stroke. *Neurology* 2010; **74**: 128–35.

159 Bhaskar S, Stanwell P, Bivard A, *et al.* The influence of initial stroke severity on mortality, overall functional outcome and in-hospital placement at 90 days following acute ischemic stroke: A tertiary hospital stroke register study. *Neurol India* 2017; **65**: 1252–9.

160 Bray JE, Coughlan K, Bladin C. Can the ABCD Score be dichotomised to identify high-risk patients with transient ischaemic attack in the emergency department? *Emerg Med J* 2007; **24**: 92–5.

161 Chandratheva A, Geraghty OC, Luengo-Fernandez R, Rothwell PM. ABCD2 score predicts severity rather than risk of early recurrent events after transient ischemic attack. *Stroke* 2010; **41**: 851–6.

162 Chandratheva A, Geraghty OC, Rothwell PM. Poor Performance of Current Prognostic Scores for Early Risk of Recurrence After Minor Stroke. *Stroke* 2011; **42**: 632–7.

163 Chatzikonstantinou A, Wolf ME, Schaefer A, Hennerici MG. Risk Prediction of Subsequent Early Stroke in Patients with Transient Ischemic Attacks. *Cerebrovasc Dis* 2013; **36**: 106–9.

164 Chen P, Liu Y, Wang Y, *et al.* A Validation of the Essen Stroke Risk Score in Outpatients with Ischemic Stroke. *J Stroke Cerebrovasc Dis* 2016; **25**: 2189–95.

165 Coutts SB, Eliasziw M, Hill MD, *et al.* An improved scoring system for identifying patients at high early risk of stroke and functional impairment after an acute transient ischemic attack or minor stroke. *Int J Stroke* 2008; **3**: 3–10.

166 Fothergill A, Christianson TJH, Brown Jr. RD, Rabinstein AA. Validation and refinement of the ABCD2 score: a population-based analysis. *Stroke* 2009; **40**: 2669–73.

167 Ghia D, Thomas P, Cordato D, *et al.* Low positive predictive value of the ABCD2 score in emergency department transient ischaemic attack diagnoses: the South Western Sydney transient ischaemic attack study. *Intern Med J* 2012; **42**: 913–8.

168 Ay H, Arsava EM, Johnston SC, *et al.* Clinical- and imaging-based prediction of stroke risk after transient ischemic attack: The CIP model. *Stroke* 2009; **40**: 181–6.

169 Johnston SC, Rothwell PM, Nguyen-Huynh MN, *et al.* Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. *Lancet* 2007; **369**: 283–92.

170 Kamouchi M, Kumagai N, Okada Y, Origasa H, Yamaguchi T, Kitazono T. Risk score for predicting recurrence in patients with ischemic stroke: The fukuoka stroke risk score for Japanese. *Cerebrovasc Dis* 2012; **34**: 351–7.

171 Kernan WN, Viscoli CM, Brass LM, *et al.* The stroke prognosis instrument II (SPI-II): A clinical prediction instrument for patients with transient ischemia and nondisabling ischemic stroke. *Stroke* 2000; **31**: 456–62.

172 Ling X, Yan SM, Shen B, Yang X. A modified Essen Stroke Risk Score for predicting recurrent ischemic stroke at one year. *Neurol Res*; **40**: 204–10.

173 Liu J, Li M, Liu J. Evaluation of the ESRS and SPI-II scales for short-term prognosis of minor stroke and transient ischemic attack. *Neurol Res* 2013; **35**: 568–72.

174 Liu Y, Wang Y, Li WA, Yan A, Wang Y. Validation of the Essen Stroke Risk Score in different subtypes of ischemic stroke. *Neurol Res* 2017; **39**: 504–8.

175 Maier IL, Bauerle M, Kermer P, Helms HJ, Buettner T. Risk prediction of very early recurrence, death and progression after acute ischaemic stroke. *Eur J Neurol* 2013; **20**: 599–604.

176 Meng X, Wang Y, Zhao X, *et al.* Validation of the Essen Stroke Risk Score and the Stroke Prognosis Instrument II in Chinese Patients. *Stroke* 2011; **42**: 3619-U448.

177 Purroy F, Jimenez Caballero PE, Gorospe A, *et al.* Prediction of early stroke recurrence in transient ischemic attack patients from the PROMAPA study: A comparison of prognostic risk scores. *Cerebrovasc Dis* 2012; **33**: 182–9.

178 Rothwell PM, Giles MF, Flossmann E, *et al.* A simple score (ABCD) to identify individuals at high early risk of stroke after transient ischaemic attack. *Lancet* 2005; **366**: 29–36.

179 Sanders LM, Srikanth VK, Psihogios H, Wong KK, Ramsay D, Phan TG. Clinical predictive value of the ABCD2 score for early risk of stroke in patients who have had transient ischaemic attack and who present to an Australian tertiary hospital. *Med J Aust* 2011; **194**: 135–8.

180 Sciolla R, Melis F, Grp S. Rapid identification of high-risk transient ischemic attacks - Prospective validation of the ABCD score. *Stroke* 2008; **39**: 297–302.

181 Sheehan OC, Merwick A, Kelly LA, *et al.* Diagnostic usefulness of the ABCD2 score to distinguish transient ischemic attack and minor ischemic stroke from noncerebrovascular events: The North Dublin TIA study. *Stroke* 2009; **40**: 3449–54.

182 Sheehan OC, Kyne L, Kelly LA, *et al.* Population-based study of ABCD2 score, carotid stenosis, and atrial fibrillation for early stroke prediction after transient ischemic attack: the North Dublin TIA study. *Stroke* 2010; **41**: 844–50.

183 Song B, Fang H, Zhao L, *et al.* Validation of the ABCD3-I score to predict stroke risk after transient ischemic attack. *Stroke* 2013; **44**: 1244–8.

184 Song B, Pei L, Fang H, *et al.* Validation of the RRE-90 scale to predict stroke risk after transient symptoms with infarction: A prospective cohort study. *PLoS One* 2015; **10**: 1–11.

185 Sumi S, Origasa H, Houkin K, *et al.* A modified Essen stroke risk score for predicting recurrent cardiovascular events: development and validation. *Int J Stroke* 2013; **8**: 251–7.

186 Tsivgoulis G, Stamboulis E, Sharma VK, *et al.* Multicenter external validation of the ABCD2 score in triaging TIA patients. *Neurology* 2010; **74**: 1351–7.

187 Weimar C, Goertler M, Rother J, *et al.* Predictive value of the Essen Stroke Risk Score and Ankle Brachial Index in acute ischaemic stroke patients from 85 German stroke units. *J Neurol Neurosurg Psychiatry* 2008; **79**: 1339–43.

188 Weimar C, Diener HC, Alberts MJ, *et al.* The Essen Stroke Risk Score predicts recurrent cardiovascular events A validation within the REduction of Atherothrombosis for Continued Health (REACH) registry. *Stroke* 2009; **40**: 350–4.

189 Weimar C, Benemann J, Michalski D, *et al.* Prediction of Recurrent Stroke and Vascular Death in Patients With Transient Ischemic Attack or Nondisabling Stroke A Prospective Comparison of Validated Prognostic Scores. *Stroke* 2010; **41**: 487–93.

190 Weimar C, Siebler M, Brandt T, *et al.* Vascular risk prediction in ischemic stroke patients undergoing in-patient rehabilitation - insights from the investigation of patients with ischemic stroke in neurologic rehabilitation (INSIGHT) registry. *Int J Stroke* 2013; **8**: 503–9.

191 Van Wijk I, Kappelle LJ, Van Gijn J, *et al.* Long-term survival and vascular event risk after transient ischaemic attack or minor ischaemic stroke: A cohort study. *Lancet* 2005; **365**: 2098–104.

192 Yang J, Fu JH, Chen XY, *et al.* Validation of the ABCD2 score to identify the patients with high risk of late stroke after a transient ischemic attack or minor ischemic stroke. *Stroke* 2010; **41**: 1298–300.

193 Han J, Choi YK, Leung WK, Hui MT, Leung MKW. Long term clinical outcomes of patients with ischemic stroke in primary care – a 9-year retrospective study. *BMC Fam Pract* 2021; **22**: 1–9.

194 Flach C, Muruet W, Wolfe CDA, Bhalla A, Douiri A. Risk and Secondary Prevention of Stroke Recurrence: A Population-Base Cohort Study. *Stroke* 2020; **51**: 2435–44.

195 Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: Systematic review and critical appraisal. *BMJ* 2019; **367**. DOI:10.1136/bmj.l5358.

196 Palazón-Bru A, Mares-García E, López-Bru D, *et al.* A critical appraisal of the clinical applicability and risk of bias of the predictive models for mortality and recurrence in patients with oropharyngeal cancer: Systematic review. *Head Neck* 2020; **42**: 763–73.

197 Carrillo-Larco RM, Altez-Fernandez C, Pacheco-Barrios N, *et al.* Cardiovascular Disease Prognostic Models in Latin America and the Caribbean: A Systematic Review. *Glob Heart* 2019; **14**: 81–93.

198 Carrillo-Larco RM, Aparcana-Granda DJ, Mejia JR, Barengo NC, Bernabe-Ortiz A. Risk scores for type 2 diabetes mellitus in Latin America: a systematic review of population-based studies. *Diabet Med* 2019; **36**: 1573–84.

199 Debray TPA, Damen JAAG, Snell KIE, *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017; **356**: i6460.

200 Feigin VL, Krishnamurthi R V., Parmar P, *et al.* Update on the Global Burden of Ischemic and Hemorrhagic Stroke in 1990-2013: The GBD 2013 Study. *Neuroepidemiology* 2015; **45**: 161–76.

201 Béjot Y, Bailly H, Graber M, *et al.* Impact of the Ageing Population on the Burden of Stroke: The Dijon Stroke Registry. *Neuroepidemiology* 2019; **52**: 78–85.

202 Chen CJ, Ding D, Starke RM, *et al.* Endovascular vs medical management of acute ischemic stroke. *Neurology* 2015; **85**: 1980–90.

203 Hajat C, Heuschmann PU, Coshall C, *et al.* Incidence of aetiological subtypes of stroke in a multi-ethnic population based study: The south London Stroke Register. *J Neurol Neurosurg Psychiatry* 2011; **82**: 527–33.

204 Syme PD, Byrne AW, Chen R, Devenny R, Forbes JF. Community-based stroke incidence in a Scottish population: the Scottish Borders Stroke Study. *Stroke* 2005; **36**: 1837–43.

205 Rothwell PM, Coull AJ, Silver LE, *et al.* Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (Oxford Vascular Study). *Lancet* 2005; **366**: 1773–83.

206 Chen Y, Wright N, Guo Y, *et al.* Mortality and recurrent vascular events after first incident stroke: a 9-year community-based study of 0·5 million Chinese adults. *Lancet Glob Heal* 2020; **8**: e580–90.

207 Sposato LA, Lam M, Allen B, Shariff SZ, Saposnik G. First-Ever Ischemic Stroke and Incident Major Adverse Cardiovascular Events in 93 627 Older Women and Men. *Stroke* 2020; **51**: 387–94.

208 NHS Digital. Hospital Episode Statistics (HES). NHS Digit. 2019. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics (accessed June 21, 2019).

209 Herrett E, Shah AD, Boggon R, *et al.* Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: Cohort study. *BMJ* 2013; **346**. DOI:10.1136/bmj.f2350.

210 Gho JMIH, Schmidt AF, Pasea L, *et al.* An electronic health records cohort study on heart failure following myocardial infarction in England: Incidence and predictors. *BMJ Open* 2018; **8**: e018331.

211 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015; **12**. DOI:10.1371/journal.pmed.1001885.

212 Feigin VL, Forouzanfar MH, Krishnamurthi R, *et al.* Global and regional burden of stroke during 1990-2010: Findings from the Global Burden of Disease Study 2010. *Lancet* 2014; **383**: 245–55.

213 Kivimäki M, Batty GD, Singh-Manoux A, Britton A, Brunner EJ, Shipley MJ. Validity of Cardiovascular Disease Event Ascertainment Using Linkage to UK Hospital Records. *Epidemiology* 2017; **28**: 735–9.

214 Doran T, Fullwood C, Gravelle H, *et al.* Pay-for-Performance Programs in Family Practices in the United Kingdom. *N Engl J Med* 2006; **355**: 375–84.

215 Wang Y, Rudd AG, Wolfe CDA. Age and ethnic disparities in incidence of stroke over time: the South London Stroke Register. *Stroke* 2013; **44**: 3298–304.

216 Phan HT, Blizzard CL, Reeves MJ, *et al.* Sex differences in long-term mortality after stroke in INSTRUCT (INternational STRoke oUtComes sTudy): A Meta-Analysis of Individual Participant Data. *Circ Cardiovasc Qual Outcomes* 2017; **10**: e003436.

217 Reeves MJ, Bushnell CD, Howard G, *et al.* Sex differences in stroke: epidemiology, clinical presentation, medical care, and outcomes. Lancet Neurol. 2008; **7**: 915–26.

218 Reeves MJ, Fonarow GC, Zhao X, Smith EE, Schwamm LH. Quality of care in women with ischemic stroke in the GWTG program. *Stroke* 2009; **40**: 1127–33.

219 Feigin VL, Norrving B, Mensah GA. Global Burden of Stroke. *Circ Res* 2017; **120**: 439–48.

220 Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: New opportunities, familiar challenges. *Clin Epidemiol* 2017; **9**: 245–50.

221 Andersen KK, Steding-Jessen M, Dalton SO, Olsen TS. Socioeconomic position and incidence of ischemic stroke in denmark 2003-2012. A nationwide hospital-based study. *J Am Heart Assoc* 2014; **3**. DOI:10.1161/JAHA.113.000762.

222 Bray BD, Paley L, Hoffman A, *et al.* Socioeconomic disparities in first stroke incidence, quality of care, and survival: a nationwide registry-based cohort study of 44 million adults in England. *Lancet Public Heal* 2018; **3**: e185–93.

223 Zhao D, Guallar E, Ouyang P, *et al.* Endogenous Sex Hormones and Incident Cardiovascular Disease in Post-Menopausal Women. *J Am Coll Cardiol* 2018; **71**: 2555–66.

224 Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. *Lancet* 1999; **353**: 89–92.

225 George J, Rapsomaniki E, Pujades-Rodriguez M, *et al.* How does cardiovascular disease first present in women and men? *Circulation* 2015; **132**: 1320–8.

226 Coull AJ, Lovett JK, Rothwell PM. Population based study of early risk of stroke after transient ischaemic attack or minor stroke: Implications for public education and organisation of services. *Br Med J* 2004; **328**: 326–8.

227 Conrad N, Judge A, Tran J, *et al.* Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *Lancet* 2018; **391**: 572–80.

228 Mach F, Baigent C, Catapano AL, *et al.* 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular riskThe Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis Society (EAS). *Eur Heart J* 2020; **41**: 111–88.

229 Salvadori E, Papi G, Insalata G, *et al.* Comparison between Ischemic and Hemorrhagic Strokes in Functional Outcome at Discharge from an Intensive Rehabilitation Hospital. *Diagnostics* 2020; **11**: 38.

230 Bhalla A, Wang Y, Rudd A, Wolfe CDA. Differences in outcome and predictors between ischemic and intracerebral hemorrhage: The South London Stroke Register. *Stroke* 2013; **44**: 2174–81.

231 Chiu D, Peterson L, Elkind MSV, Rosand J, Gerber LM, Silverstein MD. Comparison of Outcomes after Intracerebral Hemorrhage and Ischemic Stroke. *J Stroke Cerebrovasc Dis* 2010; **19**: 225–9.

232 Li L, Poon MTC, Samarasekera NE, *et al.* Risks of recurrent stroke and all serious vascular events after spontaneous intracerebral haemorrhage: pooled analyses of two population-based studies. *Lancet Neurol* 2021; **20**: 437–47.

233 Murthy SB, Zhang C, Diaz I, *et al.* Association Between Intracerebral Hemorrhage and Subsequent Arterial Ischemic Events in Participants From 4 Population-Based Cohort Studies. *JAMA Neurol* 2021; **78**: 809–16.

234 Shinozaki T, Mansournia MA, Matsuyama Y. On hazard ratio estimators by proportional hazards models in matched-pair cohort studies. *Emerg Themes Epidemiol* 2017; **14**. DOI:10.1186/s12982-017-0060-8.

235 Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data, First. Springer, 1997.

236 Jørgensen HS, Nakayama H, Raaschou HO, Olsen TS. Intracerebral hemorrhage versus infarction: Stroke severity, risk factors, and prognosis. *Ann Neurol* 1995; **38**: 45–50.

237 Barber M, Roditi G, Stott DJ, Langhorne P. Poor outcome in primary intracerebral haemorrhage: Results of a matched comparison. *Postgrad Med J* 2004; **80**: 89–92.

238 Ntaios G, Papavasileiou V, Michel P, Tatlisumak T, Strbian D. Predicting functional outcome and symptomatic intracranial hemorrhage in patients with acute ischemic stroke: A glimpse into the crystal ball? Stroke. 2015; **46**: 899–908.

239 Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? *JAMA* 2020; **323**: 1401–2.

240 Ntaios G, Lip GYH. Difficult situations in anticoagulation after stroke: Between Scylla and Charybdis. Curr. Opin. Neurol. 2016; **29**: 42–8.

241 Sembill JA, Kuramatsu JB, Schwab S, Huttner HB. Resumption of oral anticoagulation after spontaneous intracerebral hemorrhage. *Neurol Res Pract* 2019; **1**: 1–10.

242 Weitz JI, Chan NC. Advances in Antithrombotic Therapy. *Arterioscler Thromb Vasc Biol* 2019; **39**: 7–12.

243 Sagris D, Ntaios G, Georgiopoulos G, *et al.* Recommendations for lipid modification in patients with ischemic stroke or transient ischemic attack: A clinical guide by the Hellenic Stroke Organization and the Hellenic Atherosclerosis Society. *Int J Stroke* 2021; **16**: 738–50.

244 Sagris D, Ntaios G, Georgiopoulos G, Pateras K, Milionis H. Proprotein Convertase Subtilisin-Kexin Type 9 inhibitors and stroke prevention: A meta-analysis. Eur. J. Intern. Med. 2021; **85**: 130–2.

245 Ntaios G, Milionis H. Low-density lipoprotein cholesterol lowering for the prevention of cardiovascular outcomes in patients with ischemic stroke. *Int J Stroke* 2019; **14**: 476–82.

246 Milionis H, Ntaios G, Korompoki E, Vemmos K, Michel P. Statin-based therapy for primary and secondary prevention of ischemic stroke: A meta-analysis and critical overview. *Int J Stroke* 2019; **15**: 377–84.

247 Sanz-Cuesta BE, Saver JL. Lipid-Lowering Therapy and Hemorrhagic Stroke Risk: Comparative Meta-Analysis of Statins and PCSK9 Inhibitors. *Stroke* 2021; : 3142–50.

248 Krishnamurthi R V, Feigin VL, Forouzanfar MH, *et al.* Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. *Lancet Glob Heal* 2013; **1**: e259–81.

249 Strazzullo P, D'Elia L, Cairella G, Garbagnati F, Cappuccio FP, Scalfi L. Excess body weight and incidence of stroke: Meta-analysis of prospective studies with 2 million participants. Stroke. 2010; **41**. DOI:10.1161/STROKEAHA.109.576967.

250 Aune D, Sen A, Prasad M, *et al.* BMI and all cause mortality: Systematic review and non-linear dose-response meta-analysis of 230 cohort studies with 3.74 million deaths among 30.3 million participants. BMJ. 2016; **353**. DOI:10.1136/bmj.i2156.

251 Lopez-Jimenez F, Jacobsen SJ, Reeder GS, Weston SA, Meverden RA, Roger VL. Prevalence and secular trends of excess body weight and impact on outcomes after myocardial infarction in the community. *Chest* 2004; **125**: 1205–12.

252 Horwich TB, Fonarow GC, Hamilton MA, MacLellan WR, Woo MA, Tillisch JH. The relationship between obesity and mortality in patients with heart failure. *J Am Coll Cardiol* 2001; **38**: 789–95.

253 Romero-Corral A, Montori VM, Somers VK, *et al.* Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies. *Lancet* 2006; **368**: 666–78.

254 Wang ZJ, Zhou YJ, Galper BZ, Gao F, Yeh RW, Mauri L. Association of body mass index with mortality and cardiovascular events for patients with coronary artery disease: A systematic review and meta-analysis. Heart. 2015; **101**: 1631–8.

255 Doehner W, Schenkel J, Anker SD, Springer J, Audebert H. Overweight and obesity are associated with improved survival, functional outcome, and stroke recurrence after acute stroke or transient ischaemic attack: Observations from the tempis trial. *Eur Heart J* 2013; **34**: 268–77.

256 Choi H, Nam HS, Han E. Body mass index and clinical outcomes in patients after ischaemic stroke in South Korea: A retrospective cohort study. *BMJ Open* 2019; **9**. DOI:10.1136/bmjopen-2018-028880.

257 Dehlendorff C, Andersen KK, Olsen TS. Body mass index and death by stroke no obesity paradox. *JAMA Neurol* 2014; **71**: 978–84.

258 World Health Organization (WHO) Europe. Body mass index - BMI. https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi (accessed March 24, 2021).

259 Berrington de Gonzalez A, Hartge P, Cerhan JR, *et al.* Body-Mass Index and Mortality among 1.46 Million White Adults. *N Engl J Med* 2010; **363**: 2211–9.

260 Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth* 2019; **22**: 67–72.

261 Vemmos K, Ntaios G, Spengos K, *et al.* Association between obesity and mortality after acute first-ever stroke: The obesity-stroke paradox. *Stroke* 2011; **42**: 30–6.

262 Oesch L, Tatlisumak T, Arnold M, Sarikaya H. Obesity paradox in stroke ± Myth or reality? A systematic review. *PLoS One* 2017; **12**. DOI:10.1371/journal.pone.0171334.

263 The GBD 2015 Obesity Collaborators. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *N Engl J Med* 2017; **377**: 13–27.

264 Tobias DK. Addressing reverse causation bias in the obesity paradox is not 'one size fits all'. *Diabetes Care* 2017; **40**: 1000–1.

265 Preston SH, Stokes A. Obesity paradox: Conditioning on disease enhances biases in estimating the mortality risks of obesity. *Epidemiology* 2014; **25**: 454–61.

266 Janssen I, Katzmarzyk PT, Ross R. Body mass index is inversely related to mortality in older people after adjustment for waist circumference. *J Am Geriatr Soc* 2005; **53**: 2112–8.

267 Zhu S, Heshka S, Wang ZM, *et al.* Combination of BMI and waist circumference for identifying cardiovascular risk factors in whites. *Obes Res* 2004; **12**: 633–45.

268 Coutinho T, Goel K, Corrêa De Sá D, *et al.* Combining body mass index with measures of central obesity in the assessment of mortality in subjects with coronary disease: Role of 'normal weight central obesity'. *J Am Coll Cardiol* 2013; **61**: 553–60.

269 Doehner W, Gerstein HC, Ried J, *et al.* Obesity and weight loss are inversely related to mortality and cardiovascular outcome in prediabetes and type 2 diabetes: Data from the ORIGIN trial. *Eur Heart J* 2020; **41**: 2668–77.

270 Olsen TS, Dehlendorff C, Petersen HG, Andersen KK. Body mass index and poststroke mortality. *Neuroepidemiology* 2008; **30**: 93–100.

271 Wilding JPH, Batterham RL, Calanna S, *et al.* Once-Weekly Semaglutide in Adults with Overweight or Obesity. *N Engl J Med* 2021; **384**: 989–1002.

272 Ryan DH, Lingvay I, Colhoun HM, *et al.* Semaglutide Effects on Cardiovascular Outcomes in People With Overweight or Obesity (SELECT) rationale and design. *Am Heart J* 2020; **229**: 61–9.

273 Prosser J, MacGregor L, Lees KR, Diener HC, Hacke W, Davis S. Predictors of early cardiac morbidity and mortality after ischemic stroke. *Stroke* 2007; **38**: 2295–302.

274 Joosten SA, Hamza K, Sands S, Turton A, Berger P, Hamilton G. Phenotypes of patients with mild to moderate obstructive sleep apnoea as confirmed by cluster analysis. *Respirology* 2012; **17**: 99–107.

275 Haldar P, Pavord ID, Shaw DE, *et al.* Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008; **178**: 218–24.

276 Siroux V, Basagan X, Boudier A, *et al.* Identifying adult asthma phenotypes using a clustering approach. *Eur Respir J* 2011; **38**: 310–7.

277 Ahmad T, Pencina MJ, Schulte PJ, *et al.* Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J Am Coll Cardiol* 2014; **64**: 1765–74.

278 Verdonschot JAJ, Merlo M, Dominguez F, *et al.* Phenotypic clustering of dilated cardiomyopathy patients highlights important pathophysiological differences. *Eur Heart J* 2021; **42**: 162–74.

279 Seymour CW, Kennedy JN, Wang S, *et al.* Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *J Am Med Assoc* 2019; **321**: 2003–17.

280 Fereshtehnejad SM, Romenets SR, Anang JBM, Latreille V, Gagnon JF, Postuma RB. New clinical subtypes of Parkinson disease and their longitudinal progression a prospective cohort comparison with other phenotypes. *JAMA Neurol* 2015; **72**: 863–73.

281 Soria D, Garibaldi JM, Ambrogi F, *et al.* A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. *Comput Biol Med* 2010; **40**: 318–30.

282 Akyea RK, Vinogradova Y, Qureshi N, *et al.* Sex, Age, and Socioeconomic Differences in Nonfatal Stroke Incidence and Subsequent Major Adverse Outcomes. *Stroke* 2021; **52**: 396–405.

283 Altman N, Krzywinski M. The curse(s) of dimensionality. *Nat Methods* 2018; **15**: 399–400.

284 Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010; **36**: 1–13.

285 Tishbirani R. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B. 1996; : 267–88.

286 Foss AH, Markatou M. kamila: Clustering Mixed-Type Data in R and Hadoop. *J Stat Softw* 2018; **83**: 1–44.

287 Lundberg SM, Erion G, Chen H, *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020; **2**: 56.

288 Mons U, Müezzinler A, Gellert C, *et al.* Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium. *BMJ* 2015; **350**: 18.

289 Duncan MS, Freiberg MS, Greevy RA, Kundu S, Vasan RS, Tindle HA. Association of Smoking Cessation With Subsequent Risk of Cardiovascular Disease. *JAMA* 2019; **322**: 642–50.

290 Gallacher KI, Batty GD, McLean G, *et al.* Stroke, multimorbidity and polypharmacy in a nationally representative sample of 1,424,378 patients in Scotland: implications for treatment burden. *BMC Med* 2014; **12**: 1–9.

291 Miller AP, Huff CM, Roubin GS. Vascular disease in the older adult. *J Geriatr Cardiol* 2016; **13**: 727.

292 Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of Continuous Features. *Int Conf Mach Learn* 1995; : 194–202.

293 Hennig C, Liao TF. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *J R Stat Soc Ser C (Applied Stat* 2013; **62**: 309–69.

294 Foss A, Markatou M, Ray B, Heching A. A semiparametric method for clustering mixed data. *Mach Learn 2016 1053* 2016; **105**: 419–58.

295 Ichino M, Yaguchi H. Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Trans Syst Man Cybern* 1994; **24**: 698–708.

296 Preud'homme G, Duarte K, Dalleau K, *et al.* Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. *Sci Reports 2021 111* 2021; **11**: 1–14.

297 Mclachlan GJ. Cluster analysis and related techniques in medical research. *Stat Methods Med Res* 1992; **1**: 27–48.

298 Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Futur Healthc J* 2019; **6**: 94.

299 Morgan A, Sinnott SJ, Smeeth L, Minassian C, Quint J. Concordance in the recording of stroke across UK primary and secondary care datasets: a population-based cohort study. *BJGP Open* 2021; **5**: 1–11.

300 Brenner H, Savitz DA, Gefeller O. The effects of joint misclassification of exposure and disease on epidemiologic measures of association. *J Clin Epidemiol* 1993; **46**: 1195–202.

301 Rost NS, Bottle A, Lee JM, *et al.* Stroke severity is a crucial predictor of outcome: An international prospective validation study. *J Am Heart Assoc* 2016; **5**: 1–7.

302 Chohan SA, Venkatesh PK, How CH. Long-term complications of stroke and secondary prevention: an overview for primary care physicians. *Singapore Med J* 2019; **60**: 616–20.

303 Esenwa C, Gutierrez J. Secondary stroke prevention: challenges and solutions. *Vasc Health Risk Manag* 2015; **11**: 437–50.

304 Arana A, Margulis A V., Varas-Lorenzo C, *et al.* Validation of cardiovascular outcomes and risk factors in the Clinical Practice Research Datalink in the United Kingdom. *Pharmacoepidemiol Drug Saf* 2021; **30**: 247.

305 Bray BD, Cloud GC, James MA, *et al.* Weekly variation in health-care quality by day and time of admission: a nationwide, registry-based, prospective cohort study of acute stroke care. *Lancet* 2016; **388**: 170–7.

306 Garofolo KM, Yeatts SD, Ramakrishnan V, Jauch EC, Johnston KC, Durkalski VL. The effect of covariate adjustment for baseline severity in acute stroke clinical trials with responder analysis outcomes. *Trials* 2013; **14**. DOI:10.1186/1745-6215-14-98.

307 Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001; **29**: 1189–232.

308 Law MR, Wald NJ, Rudnicka AR. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *BMJ* 2003; **326**: 1423.

309 National Institute of Health and Care Excellence. Cardiovascular disease: risk assessment and reduction, including lipid modification. London: National Institute for Health and Care Excellence, 2016 https://www.nice.org.uk/guidance/cg181 (accessed Jan 31, 2018).

310 Hoo ZH, Candlish J, Teare D. What is an ROC curve? *Emerg Med J* 2017; **34**: 357–9.

Appendices

Appendix A Glossary

Appendix B Approval for CPRD Data

Appendix C Additional Results for Chapter 3

Appendix D Additional Results for Chapter 4

Appendix E Additional Results for Chapter 5

Appendix F Additional Results for Chapter 6

Appendix G Additional Results for Chapter 7

Appendix H Other publications during PhD studentship

Appendix A Glossary

Acceptable research quality flag: Patients are labelled as 'acceptable' for use in research by a process that identifies and excludes patients with non-continuous follow up or patients with poor data recording that raises suspicion as to the validity of that patient's record. Patient data is checked, for the following issues: an empty or invalid first registration date, empty or invalid current registration date, absence of a record for a year of birth, a first registration date prior to their birth year, a current registration date prior to their birth year, a transferred-out reason with no transferred-out date, a transferred-out date with no transferred-out reason, a transferred-out date prior to their first registration date, a transferred-out date prior to their current registration date, a current registration date prior to their first registration date, a gender other than Female/Male/Indeterminate, age of greater than 115 at end of follow up, recorded health care episodes in years before birth year, all recorded health care episodes have empty or invalid event dates, the registration status of temporary patients. If any of these conditions are true, the patient is labelled unacceptable and is not recommended for use in research.

Up-to-standard date: The overall quality of data in practices is mediated by the use of an 'up to standard' (UTS) date, which is deemed as the date at which data in the practice is considered to have continuous high-quality data fit for use in research. This is mediated by an analysis of the total data in the practice, which is refreshed every time a new collection for practice is processed into the database. It is based on two central concepts: assurance of continuity in data recording (gap analysis), and avoidance of the use of data for which transferred out and dead patients have been removed (death recording). The UTS date is set to the latest of these dates for each practice. CPRD recommend that analyses be performed on data following the practice UTS date.

Appendix B Approval for CPRD Data

ISAC EVALUATION OF PROTOCOLS FOR RESEARCH INVOLVING CPRD DATA

FEEDBACK TO APPLICANTS

CONFIDENTIAL		***by e-mail***	
PROTOCOL NO:	19_023R		
PROTOCOL TITLE:	Developing and validating a novel clinical severity index for cardiovascular disease in primary care		
APPLICANT:	Dr Ralph Kwame Akyea University of Nottingham Ralph.Akyea@nottingham.ac.uk		
APPROVED ☒	**APPROVED WITH COMMENTS (resubmission not required)** ☐	**REVISION/ RESUBMISSION REQUESTED** ☐	**REJECTED** ☐
INSTRUCTIONS: ***Protocols with an outcome of 'Approved' or 'Approved with comments' do not require resubmission to the ISAC.*** **REVIEWER COMMENTS:** **APPLICANT FEEDBACK:**			
DATE OF ISAC FEEDBACK:	18/03/19		
DATE OF APPLICANT FEEDBACK:			

For protocols approved from 01 April 2014 onwards, applicants are required to include the ISAC protocol in their journal submission with a statement in the manuscript indicating that it had been approved by the ISAC (with the reference number) and made available to the journal reviewers. If the protocol was subject to any amendments, the last amended version should be the one submitted.

Guidance on resubmitting applications, or making amendments to approved protocols, can be found on the CPRD website at *https://cprd.com/research-applications*

Appendix C Additional Results for Chapter 3

A population-based study exploring phenotypic characteristics of stroke using unsupervised data-driven cluster analysis

Appendix C.3.1 Search Strategy

Database: **Ovid MEDLINE(R)**

#	Searches
1	cardiovascular diseases/ or heart diseases/ or exp myocardial ischemia/ or vascular diseases/ or exp arteriosclerosis/ or cerebrovascular disorders/ or exp brain ischemia/ or exp stroke/
2	((cardio* or cardia* or heart* or coronary* or myocard* or pericard* or isch?em*) adj2 (disease? or event? or mortality)).tw.
3	((cerebrovasc* or cardiovasc* or vasc*) adj2 (disease? or event? or mortality)).tw.
4	(myocardial adj (infarct* or revascular* or re-vascular* or isch?emi*)).tw.
5	heart attack?.tw.
6	angina.tw.
7	(morbid* adj5 (cardio* or cardia* or heart* or coronary* or myocard* or pericard* or isch?em*)).tw.
8	(apoplexy or (brain adj2 accident*)).tw.
9	((brain* or cerebral or lacunar) adj2 infarct*).tw.
10	peripheral arter* disease*.tw.
11	(emboli* or arrhythmi* or thrombo* or atrial fibrillat* or atrial flutter* or tachycardi* or endocardi* or (sick adj sinus)).tw.
12	(stroke or strokes).tw.
13	cerebral vascular.tw.
14	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13
15	"Severity of Illness Index"/ and "Surveys and Questionnaires"/
16	*"Severity of Illness Index"/
17	((severity or multicomponent or multi-component or multidimensional or multi-dimensional or prognos*) adj2 (index* or indice* or survey* or tool* or questionnaire* or grad* or rate or rating or scale* or scor*)).tw.
18	(severity adj2 assess*).tw.
19	(((scor* or grad* or rate or rating or composite) adj2 (scale* or system*)) and severity).tw.
20	(stratif* and severity).tw.
21	15 or 16 or 17 or 18 or 19 or 20
22	14 and 21
23	validation stud*.pt.
24	22 and 23
25	decision model*.tw.

26	22 and 25
27	decision tree.tw.
28	22 and 27
29	prognostic model*.tw.
30	22 and 29
31	(predictive adj1 (value of tests or model)).tw.
32	22 and 31
33	(prediction adj1 (model or tool or rule)).tw.
34	22 and 33
35	(risk adj1 (assessment or score or engine or equation or algorithm or table or function or model or tool or rule)).tw.
36	22 and 35
37	(valid* or discriminat* or calibrat* or accuracy or reproducib*).ti.
38	22 and 37
39	(predict* and risk*).tw.
40	predicting.tw.
41	39 or 40
42	"reproducibility of results"/
43	"sensitivity and specificity"/
44	receiver operating characteristic*.tw.
45	ROC curve/
46	(validation or discrimination or calibration or validity or accuracy or reproducibility).tw.
47	42 or 43 or 44 or 45 or 46
48	41 and 47
49	22 and 48
50	24 or 26 or 28 or 30 or 32 or 34 or 36 or 38 or 49
51	exp animals/ not humans.sh.
52	50 not 51

Database: **Embase**

#	**Searches**
1	cardiovascular diseases/ or heart diseases/ or exp myocardial ischemia/ or vascular diseases/ or exp arteriosclerosis/ or cerebrovascular disorders/ or exp brain ischemia/ or exp stroke/
2	((cardio* or cardia* or heart* or coronary* or myocard* or pericard* or isch?em*) adj2 (disease? or event? or mortality)).tw.
3	((cerebrovasc* or cardiovasc* or vasc*) adj2 (disease? or event? or mortality)).tw.
4	(myocardial adj (infarct* or revascular* or re-vascular* or isch?emi*)).tw.
5	heart attack?.tw.
6	angina.tw.

7	(morbid* adj5 (cardio* or cardia* or heart* or coronary* or myocard* or pericard* or isch?em*)).tw.
8	(apoplexy or (brain adj2 accident*)).tw.
9	((brain* or cerebral or lacunar) adj2 infarct*).tw.
10	peripheral arter* disease*.tw.
11	(emboli* or arrhythmi* or thrombo* or atrial fibrillat* or atrial flutter* or tachycardi* or endocardi* or (sick adj sinus)).tw.
12	(stroke or strokes).tw.
13	cerebral vascular.tw.
14	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13
15	"Severity of Illness Index"/ and "Surveys and Questionnaires"/
16	*"Severity of Illness Index"/
17	((severity or multicomponent or multi-component or multidimensional or multi-dimensional or prognos*) adj2 (index* or indice* or survey* or tool* or questionnaire* or grad* or rate or rating or scale* or scor*)).tw.
18	(severity adj2 assess*).tw.
19	(((scor* or grad* or rate or rating or composite) adj2 (scale* or system*)) and severity).tw.
20	(stratif* and severity).tw.
21	15 or 16 or 17 or 18 or 19 or 20
22	14 and 21
23	validation study/
24	22 and 23
25	decision model*.tw.
26	22 and 25
27	decision tree.tw.
28	22 and 27
29	prognostic model*.tw.
30	22 and 29
31	(predictive adj1 (value of tests or model)).tw.
32	22 and 31
33	(prediction adj1 (model or tool or rule)).tw.
34	22 and 33
35	(risk adj1 (assessment or score or engine or equation or algorithm or table or function or model or tool or rule)).tw.
36	22 and 35
37	(valid* or discriminat* or calibrat* or accuracy or reproducib*).ti.
38	22 and 37
39	(predict* and risk*).tw.
40	predicting.tw.
41	39 or 40
42	*reproducibility/ or exp *validity/
43	"sensitivity and specificity"/

44	receiver operating characteristic/
45	ROC curve/
46	(validation or discrimination or calibration or validity or accuracy or reproducibility).tw.
47	42 or 43 or 44 or 45 or 46
48	41 and 47
49	22 and 48
50	24 or 26 or 28 or 30 or 32 or 34 or 36 or 38 or 49
51	(exp animals/ or nonhuman/) not human/
52	50 not 51
53	Conference*.pt.
54	52 not 53

Database: **PsycINFO**

#	**Searches**
1	exp cardiovascular disorders/
2	((cardio* or cardia* or heart* or coronary* or myocard* or pericard* or isch?em*) adj2 (disease? or event? or mortality)).tw.
3	((cerebrovasc* or cardiovasc* or vasc*) adj2 (disease? or event? or mortality)).tw.
4	(myocardial adj (infarct* or revascular* or re-vascular* or isch?emi*)).tw.
5	heart attack?.tw.
6	angina.tw.
7	(morbid* adj5 (cardio* or cardia* or heart* or coronary* or myocard* or pericard* or isch?em*)).tw.
8	(apoplexy or (brain adj2 accident*)).tw.
9	((brain* or cerebral or lacunar) adj2 infarct*).tw.
10	peripheral arter* disease*.tw.
11	(emboli* or arrhythmi* or thrombo* or atrial fibrillat* or atrial flutter* or tachycardi* or endocardi* or (sick adj sinus)).tw.
12	(stroke or strokes).tw.
13	cerebral vascular.tw.
14	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13
15	((severity or multicomponent or multi-component or multidimensional or multi-dimensional or prognos*) adj2 (index* or indice* or survey* or tool* or questionnaire* or grad* or rate or rating or scale* or scor*)).tw.
16	(severity adj2 assess*).tw.
17	(((scor* or grad* or rate or rating or composite) adj2 (scale* or system*)) and severity).tw.
18	(stratif* and severity).tw.
19	15 or 16 or 17 or 18
20	14 and 19

21	decision model*.tw.
22	20 and 21
23	decision tree.tw.
24	20 and 23
25	prognostic model*.tw.
26	20 and 25
27	(predictive adj1 (value of tests or model)).tw.
28	20 and 27
29	(prediction adj1 (model or tool or rule)).tw.
30	20 and 29
31	(risk adj1 (assessment or score or engine or equation or algorithm or table or function or model or tool or rule)).tw.
32	20 and 31
33	(valid* or discriminat* or calibrat* or accuracy or reproducib*).ti.
34	20 and 33
35	(predict* and risk*).tw.
36	predicting.tw.
37	35 or 36
38	exp test reliability/
39	exp test validity/
40	receiver operating characteristic*.tw.
41	(validation or discrimination or calibration or validity or accuracy or reproducibility).tw.
42	38 or 39 or 40 or 41
43	37 and 42
44	20 and 43
45	22 or 24 or 26 or 28 or 30 or 32 or 34 or 44

Database: **Web of Science**

Set	Save History
# 44	#42 OR #38 OR #37 OR #36 OR #35 OR #34 OR #33 Refined by: DOCUMENT TYPES: (ARTICLE) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 43	#42 OR #38 OR #37 OR #36 OR #35 OR #34 OR #33 Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 42	#41 AND #32 Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 41	#40 AND #39 Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 40	#29 OR #27 OR #26 OR #25 OR #17 Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 39	#21 OR #20 Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 38 #32 AND #23
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 37 #32 AND #22
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 36 #32 AND #19
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 35 #32 AND #18
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 34 #32 AND #24
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 33 #32 AND #28
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 32 #31 AND #30
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 31 #16 OR #15 OR #14 OR #13 OR #12 OR #11 OR #10 OR #9 OR #8 OR #7 OR #6 OR #5
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 30 #4 OR #3 OR #2 OR #1
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 29 TS=("reproducibility of results")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 28 TS=("validation stud*")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 27 TS=(validation OR discrimination OR calibration OR validity OR accuracy OR reproducibility)
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 26 TS=("sensitivity and specificity")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 25 TS=("ROC curve")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 24 TS=("decision tree")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 23 TS=(risk NEXT (assessment OR score OR engine OR equation OR algorithm OR table OR function OR calculator OR calculation OR function OR model OR tool OR rule))
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 22 TS=(predict* NEXT ("value of tests" OR model OR tool OR rule))
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 21 TS=("predicting")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 20 TS=(predict* and risk*)
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 19 TS=("prognostic model*")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 18 TS=("decision model*")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 17 TS=("receiver operating characteristic*")
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 16 TS=((cardio* or cardia* or heart* or coronary* or myocard* or pericard* or ischem* or ischaemi*) NEAR/2 (disease* or event* or mortality))
Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

# 15	TS=((cerebrovasc* or cardiovasc* or vasc*) NEAR/2 (disease* or event* or mortality)) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 14	TS=(myocardial NEAR/2 (infarct* or revascular* or re-vascular* or ischem* or ischaemi*)) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 13	TS=("heart attack*") Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 12	TOPIC: (angina) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 11	TS=(morbid* NEAR/5 (cardio* or cardia* or heart* or coronary* or myocard* or pericard* or ischem* or ischaemi*)) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 10	TOPIC: (apoplexy or (brain NEAR/2 accident*)) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 9	TOPIC: ((brain* or cerebral or lacunar) NEAR/2 infarct*) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 8	TS=("peripheral arter* disease*") Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 7	TS=(emboli* or arrhythmi* or thrombo* or atrial fibrillat* or atrial flutter* or tachycardi* or endocardi* or (sick NEXT sinus)) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 6	TOPIC: (stroke or strokes) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 5	TS=("cerebral vascular") Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 4	TOPIC: ((severity or multicomponent or multi-component or multidimensional or multi-dimensional or prognos*) NEAR/2 (index* or indice* or survey* or tool* or questionnaire* or grad* or rate or rating or scale* or scor*)) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 3	TOPIC: (severity NEAR/2 assess*) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 2	TS=(((scor* or grad* or rate or rating or composite) NEAR/2 (scale* or system*)) and severity) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years
# 1	TOPIC: (stratif* and severity) Indexes=SCI-EXPANDED, CPCI-S Timespan=All years

Appendix C.3.2 Characteristics of included validation studies

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study populations	Outcome(s) predicted	Follow-up for outcome
Andersen 2015	CHA2DS2VASc score Essen Stroke Risk Score	2003 - 2012	Prospective	Linked Danish Stroke Registry: 42,182	Linked Danish Stroke Registry: • Denmark • Hospital setting • Age: 70.1 years (19.3) • Female: 45.7%	Recurrent ischemic stroke, death, or a cardiovascular event.	1 & 5 years
Andersen 2017	CHA2DS2VASc score Essen Stroke Risk Score	2005 - 2012	Retrospective	Linked Danish Stroke Registry: 832	Linked Danish Stroke Registry: • Denmark • Hospital setting • Age: 59.6 years (13.9) • Female: 349 (42%)	Recurrent ischemic stroke; Death; Cardiovascular events	3.3 years (SD 2.1)
Arsava 2011	Recurrence Risk Estimator (RRE) score ABCD2 score	2003 - 2009	Retrospective	302 admitted 257 with complete follow-up hence analysed	Patient with TIA: • USA • Hospital setting • Age: 67 years (55-76) • Female: 124 (48.2%)	Recurrent ischemic stroke	7 days
Arsava 2016	Recurrence Risk Estimator (RRE) score	USA cohort: 2009-2011 South Korea & Brazil: 2007 - 2011	Retrospective cohorts for SK & Brazil Prospective for the US	1,468 Discrimination analysis: 1,331	Patients with ischaemic stroke: • USA, Brazil, South Korea • Hospital setting • Age: 69 years (58-79) • Female: 633 (43.1%)	Recurrent ischaemic stroke	90 days
Asimos 2010	ABCD2 Score	From 2005 for 35 months	Prospective	North Carolina Collaborative Stroke Registry: 1,667	Patient with TIA: • USA • Hospital setting • Age: 67.4 years (15.1) • Males: 754 (45.2%)	Ischaemic stroke	7 days
Bray 2007	ABCD Score	2004	Cohort	102 Follow-up for 98	Patients with TIA: • Australia • Hospital (ER) setting • Age: 73 years (14.5) • Female: 51 (50%)	Stroke	90 days

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study populations	Outcome(s) predicted	Follow-up for outcome
Chandratheva 2010	ABCD2 score	2002 – 2007	Not explicitly stated but is prospective cohort	The Oxford Vascular (OXVASC) Study: 500	Patients with TIA and stroke: • England • Primary care setting • Age: 72.5 years (12.7) • Male: 219 (43.8%)	recurrent Stroke (minor and major), recurrent TIA	30 days
Chandratheva 2011	ABCD2 Score	2002 – 2007	Prospective	The Oxford Vascular (OXVASC) Study: 520 minor strokes	Patients with TIA and stroke: • England • Primary care setting • Age: 73.4 years (12.3) • Male: 241 (54%)	Recurrent stroke	7 days 90 days
Chatzikonstantinou 2013	ABCD2 score ABCD3-I score	Not reported	Cohort	235	Patients with TIA and stroke: • Germany • Stroke centre setting • Age: 66.1 years (13.9) • Male: 130 (55.3%)	Early stroke	Not reported
Chen 2016	Essen Stroke Risk Score (ESRS)	2010 – 2012	Prospective	Registry of Outpatients with Ischemic Stroke in Urban China (ROOTS): 3,316	Patients with ischaemic stroke: • China • Hospital setting • Age: 63.8 years (12) • Male: 130 (68.1%)	Recurrent stroke Combined vascular events - *recurrent stroke, MI, vascular death, angina pectoris or TIA*	3 months 6 months 12 months
Coutts 2008	ABCD2 score ABCD2 + MRI	Not reported	Prospective	The VISION cohort study: 180	Patients with TIA or minor stroke: • Canada • Hospital setting • Age: 65.6 years (13.8) • Female: 69 (38.3%)	Recurrent stroke Functional impairment	90 days
Fothergill 2009	ABCD score ABCD2 score	1985 - 1994	Cohort	Rochester Stroke and Transient Ischemic Attack Registry: 284	Patients with TIA: • USA • Hospital setting • Age: 71.9 years (13.6) • Female: 158 (56%)	Stroke and death	1 year – 7, 30 & 365 days

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study populations	Outcome(s) predicted	Follow-up for outcome
Ghia 2012	ABCD2 score	2004 - 2006	Cohort	The Southwestern Sydney Transient Ischaemic Attack Study: 827 with TIA 789 analysed	Patients with TIA: • Australia • Hospital (ER) setting • Age: 69.6 years (14.6) • Male: 50.2%	Stroke	1 year - 30, 90 & 365 days
Johnston 2007	The California score ABCD score ABCD2 score	**Derivation:** California emergency dept: 1997 – 1998 Oxford population-based: 1981 – 1986 **Validation:** California emergency dept: 1998 – 1999 California clinic: 1998 – 1999 Oxford population-based: 2002 – 2005 Oxford clinic: 2002 – 2005	Cohort	**2 derivation cohorts** California emergency dept: 1,707 Oxford population-based: 209 **4 validation cohorts** California emergency dept: 1,069 California clinic: 962 Oxford population-based: 547 Oxford clinic: 315 2,893	Patients with TIA: • USA, UK • Hospital, specialist clinics and primary care settings • Age *(those 60 years & over)*: – California ED (dev.): 1325 (78%) – Oxford P-B: 167 (80%) – California ED (val.): 872 (80%) – California clinic: 722 (75%) – Oxford P-B (val.): 411 (75%) – Oxford clinic: 208 (66%) • Female: – California ED (dev.): 899 (53%) – Oxford P-B: 97 (46%) – California ED (val.): 559 (52%) – California clinic: 507 (53%) – Oxford P-B (val.): 300 (55%) – Oxford clinic: 171 (54%)	Stroke	2 days 7 days 90 days

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study populations	Outcome(s) predicted	Follow-up for outcome
Liu 2013	Essen stroke risk score Stroke prognosis instrument II (SPI-II)	2009 - 2011	Prospective	CHANCE database: 167	Patients with minor stroke and TIA: • China • Hospital setting • Age: 61.1 years (10.8) • Female: 48 (28.7%)	Cardiovascular or cerebrovascular ischaemic events	90 days
Liu 2017	Essen Stroke Risk Score	2011 - 2014	Prospective	Blood pressure and clinical Outcome in Stroke Survivors (BOSS) Nationwide Registry: Total with ischaemic stroke: 1,699 *Sub-types:* Large-artery atherosclerosis (LAA): 972 Small-artery occlusion (SAO): 635	Patients with stroke: • China • Hospital setting • Age: – IS: 62.0 years (10.7) – LAA: 62.6 years (10.4) – SAO: 61.5 years (10.5) • Female: – IS: 523 (31.8%) – LAA: 287 (29.5%) – SAO: 206 (32.4%)	Recurrent stroke	1 year
Maier 2013	Essen Stroke Risk Score ABCD2 score Recurrence Risk Estimator (RRE) score	2007 - 2011	Cohort	1,727	Patients with ischaemic stroke: • Germany • Hospital setting • Age: 71 years (13) • Male: 964 (55.8%)	Recurrence early Recurrence in-patient Death early Death in-patient Progressive stroke	7 days
Meng 2011	Essen Stroke Risk Score Stroke Prognostic Instrument II	2007 - 2008	Prospective	The China National Stroke Register (CNSR): 11,384 TIA: 1,061 Ischaemic stroke: 10,323 Analysed: 9,152	Patients with ischaemic stroke and TIA: • China • Hospital setting • Age: Not reported • Males: – With follow-up: 7,222 (63.4%)	Recurrent stroke and combined vascular event.	1 year

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study populations	Outcome(s) predicted	Follow-up for outcome
Purroy 2012	ABCD score ABCD2 score ABCD2I score ABCDI ABCD3 score ESRS SPI-II California scale	2008 - 2009	Prospective	PROMAPA Study: 1,255 enrolled Analysis: 1,137	Patients with TIA: • Spain • Hospital setting • Age: 68.6 years (13.1) • Male: 674 (59.3%)	Recurrent stroke	7 days 90 days
Sanders 2011	ABCD2 score	2004 - 2007	Prospective	512 289 at 90 days	Patients with TIA: • Australia • Hospital setting • Age: Range for cut-offs, 61.1±12.7 to 74.3±9.4 • Male: 175	Stroke	2 days 90 days
Sciolla 2008	ABCD score ABCDI score	May - Oct 2006	Prospective	SINPAC group: 274	Patients with TIA: • Italy • Hospital setting • Age: 71.5 years (10.5) • Male: 169 (61.7%)	Stroke	7 days 1 month
Sheehan 2009	ABCD2 score	2005 - 2007	Prospective	North Dublin TIA Study *(sub-study of North Dublin Population Stroke Study)*: 594	Patients with TIA: • Ireland • Community & hospital settings • Age: 65±12 to 70±13 • Female: 329 (55.4%)	TIA and minor ischemic stroke (MIS) and non-cerebrovascular event	Not reported **ABCD2 score at referral and then diagnosis made at clinic*

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study populations	Outcome(s) predicted	Follow-up for outcome
Sheehan 2010	ABCD2 score	2005 - 2008	Prospective	North Dublin TIA Study: Specialist-confirmed TIA: 443 Non-specialist referrals for TIA: 700	Patients with TIA: • Ireland • Community & secondary care settings • Age (years): – Specialist-confirmed: 70 (13) – Non-specialist referrals: 68 (13) • Female: – Specialist-confirmed: 230 (53.1) – Non-specialist referrals: 387 (55.3)	Stroke	7 days 28 days 90 days
Song 2013	ABCD2 score ABCD3-I score	2010 - 2011	Prospective	Zhengzhou University Cohort Study: 239	Patients with TIA: • China • Hospital setting • Age: 57.4 years (13.32) • Female: 96 (40.2%)	Stroke	90 days
Song 2015	RRE-90 score ABCD2 score	2010 - 2014	Prospective	Zhengzhou University Cohort Study: 221	Patients with TIA and RRE completed: • China • Hospital setting • Age: 57.48 years (12.72) • Female: 87 (39.4%)	Ischemic stroke	90 days
Tsivgoulis 2010	ABCD2 score	2008 - 2009	Prospective	148	Patients with TIA: • Greece, Singapore • Hospital setting • Age: 60 years (14) • Female: 55	Stroke	7 days 90 days

Lead author and Year	Model(s)	Study period	Study design	Number of study participants	Study populations	Outcome(s) predicted	Follow-up for outcome
Weimar 2008	Essen Stroke Risk Score Ankle Brachial Index (ABI)	2005	Prospective	852	Patients with ischaemic stroke or TIA: • Germany • Hospital setting • Age: 67.1 years (12.4) • Female: not reported	Recurrent cerebrovascular events - stroke	17.5 months
Weimer 2009	Essen Stroke Risk Score	2003 - 2004	Prospective	Reduction of Atherothrombosis for Continued Health (REACH) Registry: 15,605	Patients with stroke or TIA: • International • Community and hospital settings • Age: 68.9 years (10.1) • Female: Not reported	Major cardiovascular events *(cardiovascular death, myocardial infarction, or stroke)* Fatal and non-fatal stroke	1 year
Weimar 2010	Essen Stroke Risk Score Hankey et al LiLAC score SPI-II score	2005 - 2006	Prospective	2,381 *1,897 followed up for > 6 months*	Patients with stroke or TIA: • Germany • Hospital setting • Age *(those with follow-up)*: 67.7 years (12.3) • Male *(those with follow-up)*: 363 (54.1%)	Fatal and non-fatal stroke or cardiovascular death	Median: 1 year
Weimar 2012	Essen Stroke Risk Score Stroke Prognosis Instrument II	May - Sept 2008	Prospective	INSIGHT Registry: 1,163 *(856 with follow up)*	Patients with ischaemic stroke: • Germany • Neuro rehab unit setting • Age: 66.3 years (12.3) • Male: 662 (57.5%)	Non-fatal MI, recurrent stroke, or transitory ischemic attack as well as a cause of death	1 year
Yang 2010	ABCD2 score	2004 - 2005	Prospective	490	Patients with TIA or minor stroke: • Hong Kong, China • Hospital setting • Age: 66.3 years (13.1) • Female: 216 (44%)	Stroke and death	Mean: 40.5 months (SD 10.7)

Appendix C.3.3 Predictive accuracy of prognostic models being validated

Lead author & Year	Outcome(s)	Number of study participants	Participants with outcome(s)	Model(s)	AUC or C-statistics with 95% CI	Calibration
Andersen 2015	Recurrent ischemic stroke	42,182	1 year: 1,312	CHA2DS2VASc score	0.52 (0.51 - 0.53)	No information
				Essen stroke score	0.54 (0.53-0.55)	
			5 years: 2,858	CHA2DS2VASc score	0.54 (0.53 - 0.55)	
				Essen stroke score	0.56 (0.55-0.57)	
	Death *(all-cause)*, or					
	Cardiovascular event		1 year: 2,382	CHA2DS2VASc score	0.53 (0.52-0.54)	
				Essen stroke score	0.55 (0.54-0.56)	
			5 years: 5,032	CHA2DS2VASc score	0.55 (0.54-0.56)	
				Essen stroke score	0.57 (0.57-0.58)	
Andersen 2017	• Recurrent ischemic stroke • Death • Cardiovascular events defined as either ischemic stroke, TIA, MI or arterial thromboembolism	832	Recurrent ischaemic stroke: 55	CHA2DS2VASc score	0.59 (0.51–0.65)	No information
				Essen Stroke Score	0.60 (0.53–0.68)	
				CHA2DS2VASc score + DWMH score ≥2	0.62 (0.54–0.70)	
				Essen Stroke Score + DWMH score ≥2	0.63 (0.56–0.71)	
				DWMH score	0.65 (0.58–0.73)	
				PVH score	0.62 (0.52–0.68)	
				Total Fazekas score	0.65 (0.58–0.73)	
Arsava 2011	Recurrent ischemic stroke	Admitted: 302 Analysed with complete follow-up: 257	24	Recurrence Risk Estimator (RRE) score	0.85 (0.78-0.92)	No information
				RRE score *(all 302)*	0.86 (0.80-0.93)	
				ABCD2 score	0.57 (0.45-0.69)	
Arsava 2016	Recurrent ischaemic stroke	1,468 Discrimination analysis: 1,331	59	Recurrence Risk Estimator (RRE) score	90-day recurrence: 0.76 (0.70-0.82) Clinical model: 0.65 (0.59-0.71)	Calibration slope: 0.61 Hosmer-Lemeshow test: p = 0.008

Lead author & Year	**Outcome(s)**	**Number of study participants**	**Participants with outcome(s)**	**Model(s)**	**AUC or C-statistics** with 95% CI	**Calibration**
Asimos 2010	Ischaemic stroke	1,667 Complete case: 1,054	373 (23%)	ABCD2 Score	Complete case: 0.59 (0.56-0.62) Imputed: 0.58 (0.56-0.61)	No information
Bray 2007	Stroke	102 Follow-up for 98	7	ABCD Score	No information	No information
Chandratheva 2010	Recurrent stroke (major and minor stroke) Recurrent TIA	500	Recurrent TIA: 55 (11.0%) Recurrent stroke: 50 *(28 major strokes versus 22 minor strokes)*	ABCD2 score *(recurrent stroke within 7 days)*	0.71 (0.63-0.79)	No information
				ABCD2 score *(major recurrent stroke within 7 days)*	0.80 (0.72-0.87)	
				ABCD2 score *(minor recurrent stroke within 7 days)*	0.57 (0.43-0.71)	
				ABCD2 score *(recurrent TIA within 7 days)*	0.37 (0.29-0.44)	
Chandratheva 2011	Recurrent stroke	Minor strokes: 520	Recurrent strokes within 90-days: 142 • Within 7 days: 81 • Within 30 days: 111	ABCD2 Score *(within 90 days)*	0.60 (0.52-0.67)	No information
				ABCD2 Score *(within 7 days)*	0.57 (0.47-0.68)	
				ABCD2 Score *(within 90 days from first call for medical attention)*	0.62 (0.54-0.70)	
				ABCD2 Score *(within 7 days from first call for medical attention)*	0.64 (0.53-0.74)	
				Essen Stroke Score *(within 90 days)*	0.50 (0.42-0.59)	
				Essen Stroke Score *(within 7 days)*	0.49 (0.35-0.62)	
				SPI-II *(within 90 days)*	0.48 (0.39-0.60)	
				SPI-II *(within 7 days)*	0.50 (0.37-0.64)	

Lead author & Year	**Outcome(s)**	**Number of study participants**	**Participants with outcome(s)**	**Model(s)**	**AUC or C-statistics** with 95% CI	**Calibration**
Chatzikonstantinou 2013	Early stroke	235	17	ABCD2 score ABCD3-I score	No information	No information
Chen 2016	Recurrent stroke	3,316	1-year cumulative occurrence rate for recurrent stroke: 82 2.47% (1.97-3.06%)	Essen Stroke Risk Score *(at 3 months)*	0.6303 (0.5308-0.7298)	No information
				Essen Stroke Risk Score *(at 6 months)*	0.6156 (0.5350-0.6952)	
				Essen Stroke Risk Score *(at 12 months)*	0.6283 (0.5683-0.6883)	
	Composite vascular events: including recurrent stroke, MI, vascular death, angina pectoris and TIA		1-year cumulative occurrence rate for recurrent stroke: 143 4.32% (3.65-5.06%)	Essen Stroke Risk Score *(at 3 months)*	0.6079 (0.5310-0.6848)	
				Essen Stroke Risk Score *(at 6 months)*	0.6256 (0.5661-0.6851)	
				Essen Stroke Risk Score *(at 12 months)*	0.6295 (0.5836-0.6754)	
Coutts 2008	**Recurrent stroke** Functional impairment	180	20	ABCD2	0.78 (0.68-0.87)	No information
				MRI	0.84 (0.72-0.92)	
				ABCD2+MRI	0.88 (0.79-0.94)	
				All factors	0.90 (0.81-0.95)	
Fothergill 2009	**Stroke** and death	284	Within 7 days: 36 Within 30 days: 41 Within 365 days: 64	ABCD score ABCD2 score	Stroke (ABCD2): 7 days: 0.654 30 days: 0.653 356 days: 0.635	No information
Ghia 2012	Stroke	With TIA: 827 Analysed: 789	Within 2 days: 3 Within 30 days: 7 Within 90 days: 15 Within 1 year: 19	ABCD2 score	No information	No information

Lead author & Year	**Outcome(s)**	**Number of study participants**	**Participants with outcome(s)**	**Model(s)**	**AUC or C-statistics** with 95% CI	**Calibration**
Johnston 2007	Stroke	2,893	At 2 days: 189 (3·9%)	The California Score	*48 AUCs reported for days 2, 7 & 30 for all 3 models, for all the 5 cohorts:* Range from 0.60 (0.52-0.69) to 0.75 (0.66-0.85)	No information
				ABCD score	Range from 0.62 (0.55-0.70) to 0.76 (0.68-0.83)	
				ABCD2 score	Range from 0.62 (0.54-0.69) to 0.79 (0.68-0.90)	
			At 7 days: 267 (5·5%)	The California Score	Range from 0.60 (0.54-0.67) to 0.79 (0.72-0.87)	
				ABCD score	Range from 0.64 (0.57-0.70) to 0.81 (0.73-0.89)	
				ABCD2 score	Range from 0.63 (0.57-0.69) to 0.83 (0.75-0.91)	
			Within 90-days: 442 (9·2%)	The California Score	Range from 0.61 (0.56-0.67) to 0.73 (0.64-0.82)	
				ABCD score	Range from 0.63 (0.58-0.68) to 0.77 (0.67-0.86)	
				ABCD2 score	Range from 0.64 (0.58-0.69) to 0.75 (0.67-0.84)	
Liu 2013	Cardiovascular or cerebrovascular ischaemic events	167	Recurrent ischaemic event: 21 Cerebral ischaemic event: 20	Essen stroke risk score	0.677 (0.557-0.797)	No information
				Stroke prognosis Instrument II (SPI-II)	0.553 (0.413-0.694)	

Lead author & Year	**Outcome(s)**	**Number of study participants**	**Participants with outcome(s)**	**Model(s)**	**AUC or C-statistics** with 95% CI	**Calibration**
Liu 2017	Recurrent stroke	Total: 1,699 By sub-type: IS 1,699 LAA 972 SAO 635	Recurrent stroke: 186 By sub-type: IS 97, LAA 60, SAO 29 Composite vascular events: 211 By sub-type: IS 110, LAA 69, SAO 32	Essen stroke risk score	Recurrent stroke: IS: 0.581 (0.524-0.638) LAA 0.609 (0.538-0.680), SAO 0.563 (0.470-0.655) Composite vascular events: IS: 0.585 (0.531-0.639) LAA: 0.607 (0.539-0.675) SAO 0.579 (0.490-0.668)	Hosmer-Lemeshow: p = 0.35 p = 0.23 p = 0.52 p = 0.30 p = 0.14 p = 0.61
Maier 2013	Recurrence early Death early	1,727	Recurrence early: 56 (3.2%) Death early: 40 (2.3%)	Essen Stroke Score	0.50 (0.42–0.58)	No information
				ABCD2 score	0.60 (0.53–0.67)	
				Recurrence Risk Estimator (RRE) score	0.65 (0.58–0.73)	
	Progressive stroke		125 (7.2%)			
	Recurrence in-patient Death in-patient		Recurrence in-patient (n, %) 39 (2.3) Death in-patient (n, %) 30 (1.7)	Essen Stroke Score	0.59 (0.50–0.68)	
				ABCD2 score	0.60 (0.52–0.69)	
				Recurrence Risk Estimator (RRE) score	0.72 (0.64–0.80)	
Meng 2011	Recurrent stroke	11,384 TIA: 1,061 or Ischaemic stroke: 10,323 Analysed: 9,152	Not reported	Essen Stroke Score	0.59 (0.58-0.60)	No information
				Stroke Prognostic Instrument II	0.59 (0.58 0.61)	
	Combined vascular event			Essen Stroke Score	0.60 (0.59-0.61)	
				Stroke Prognostic Instrument II	0.60 (0.58-0.61)	

Lead author & Year	**Outcome(s)**	**Number of study participants**	**Participants with outcome(s)**	**Model(s)**	**AUC or C-statistics** with 95% CI	**Calibration**
Purroy 2012	Stroke recurrence	Enrolled: 1,255 Analysis: 1,137	Within 7 days: 29 (2.6%)	ABCD score	0.57 (0.46–0.68)	No information
				ABCD2 score	0.56 (0.45–0.66)	
				ABCD2I score	0.56 (0.45–0.67)	
				ABCDI	0.56 (0.44–0.67)	
				ABCD3 score	0.66 (0.57–0.81)	
				ABCD3V	0.69 (0.57–0.81)	
				Essen stroke score	0.60 (0.51–0.70)	
				Stroke prognosis Instrument II (SPI-II)	0.50 (0.41–0.59)	
				California Risk scale	0.52 (0.42–0.63)	
			Within 90 days: 43 (3.8%)	ABCD score	0.55 (0.46–0.64)	
				ABCD2 score	0.55 (0.46–0.64)	
				ABCD2I score	0.55 (0.44–0.65)	
				ABCDI	0.56 (0.45–0.67)	
				ABCD3 score	0.61 (0.52–0.70)	
				ABCD3V	0.63 (0.54–0.73)	
				Essen stroke score	0.58 (0.49–0.66)	
				Stroke prognosis Instrument II (SPI-II)	0.51 (0.43–0.60)	
				California Risk scale	0.54 (0.45–0.63)	
Sanders 2011	Stroke	512 With follow-up at 90 days: 289	Within 2 days: 4/292	ABCD2 score	0.80 (0.68-0.91)	No information
			Within 90 days: 7/289		0.62 (0.4-0.83)	
Sciolla 2008	Stroke	274	7 days: 10 (3.6%)	ABCD score	0.75 (0.63-0.88);	No information
				ABCDI score	0.78 (0.65-0.91)	
			30 days: 15 (5.5%)	ABCD score	0.76 (0.66-0.86)	
				ABCDI score	0.79 (0.69-0.89)	

Lead author & Year	**Outcome(s)**	**Number of study participants**	**Participants with outcome(s)**	**Model(s)**	**AUC or C-statistics** with 95% CI	**Calibration**
Sheehan 2009	TIA and minor ischemic stroke (MIS) and non-cerebrovascular event	594	TIA: 292 (49.2%) MIS: 45 (7.6%) TIA+MIS: 337 (56.7%) Non-cerebrovascular: 257 (43.3%)	ABCD2 score	TIA: 0.68 (0.64-0.72) TIA + MIS: 0.7 (0.66-0.74) MIS: 0.81 (0.75-0.87)	No information
Sheehan 2010	Stroke	Confirmed TIA: 443 Non-specialist suspected TIA: 700	Recurrent (ischaemic) stroke: 7 days: 15 (3.4%) 28 days: 24 (5.4%) 90 days: 33 (7.5%)	ABCD2 score	Confirmed TIA: 7 days: 0.49 (0.35-0.63) 28 days: 0.55 (0.43-0.66) 90 days: 0.55 (0.45-0.64) Non-specialist suspected TIA 7 days: 0.56 (0.42-0.70) 28 days: 0.61 (0.50-0.72) 90 days: 0.61 (0.52-0.71)	No information
Song 2013	Stroke	239	29 (12.13%)	ABCD3-I score	0.825 (0.752-0.898)	No information
				ABCD2 score	0.694 (0.601-0.786)	
Song 2015	Ischemic stroke	With RRE completed: 221	46 (20.81%)	Recurrence Risk Estimator (RRE)-90 score	0.681 (0.592-0.771)	No information
				ABCD2 score	0.546 (0.454-0.638)	
Tsivgoulis 2010	Stroke	148	7 days: n=12 (8%)	ABCD2 score	0.72 (0.57-0.88)	No information
			90 days: n=24 (16%)		0.75 (0.65-0.86)	
Weimar 2008	Recurrent cerebrovascular events - stroke	852	Stroke: 41 (5.6%) TIA: 15 (2.1%) Death (all-cause): 52 (7.1%) CV death: 33 (4.5%)	Essen Stroke Risk Score (ESRS) Ankle Brachial Index (ABI)	ESRS for stroke: 0.56 ESRS (composite vascular stroke or cv death): 0.61 (0.54-0.59)	No information
Weimer 2009	Composite major CV events: CV death, MI, and stroke)	15,605	6.05%	Essen stroke risk score	0.60 (0.58-0.62)	No information
	Fatal and non-fatal stroke		4.01%		0.56 (0.53-0.58)	

Lead author & Year	**Outcome(s)**	**Number of study participants**	**Participants with outcome(s)**	**Model(s)**	**AUC or C-statistics** with 95% CI	**Calibration**
Weimar 2010	Fatal and nonfatal stroke	2,381 With follow up > 6 months: 1,897	Recurrent cerebral stroke: 107 (5.6%) • Fatal or major strokes with persisting disability: 42	Essen stroke score	0.62 (0.57-0.67)	No information
				Hankey et al	0.62 (0.57-0.67)	
				LiLAC score	0.64 (0.59-0.69)	
				Stroke Prognostic Instrument II	0.65 (0.60-0.70)	
	Stroke or cardiovascular (CV) death		Deaths, 75 (4.0%): • cerebral stroke: 13 • Other CV death: 28 • Non-CV death: 27 • Unknown cause: 7	Essen stroke score	0.65 (0.60-0.69)	
				Hankey et al	0.64 (0.60-0.69)	
				LiLAC score	0.65 (0.61-0.70)	
				Stroke Prognostic Instrument II	0.66 (0.61-0.70)	
Weimar 2012	Nonfatal MI, recurrent stroke, or transitory ischemic attack as well as a cause of death	1,163 With complete follow up: 846	Recurrent stroke: 6.7%	Essen Stroke Score	0.62 (0.59-0.65)	No information
				Stroke Prognostic Instrument II	0.56 (0.53-0.60)	
			Combined vascular events: 10.9%	Essen Stroke Score	0.59 (0.56-0.63)	
				Stroke Prognostic Instrument II	0.60 (0.57-0.64)	
Yang 2010	Stroke and death	490	Stroke: 76 (15.5%) Death: 62 (12.7%)	ABCD2 score	Further stroke: 0.65 (0.58 - 0.71) Death: 0.59 (0.52 - 0.67)	No information

Appendix D Additional Results from Chapter 4

Sex, age, and socioeconomic differences in non-fatal stroke incidence and subsequent major adverse outcomes

Appendix D.4.1 Incidence of stroke presented by year and sex (1998 – 2017)

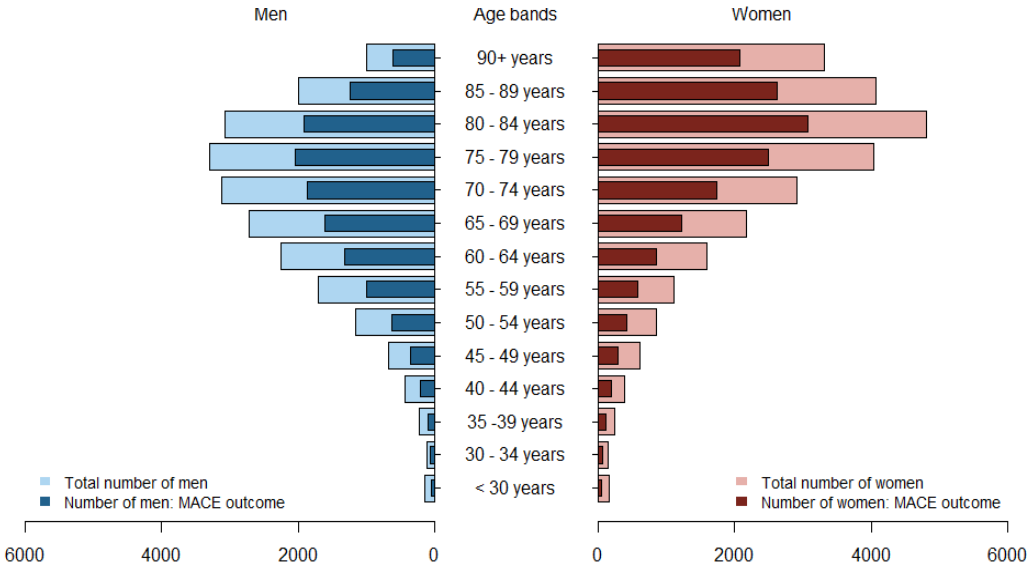
	Stroke events		**100,000 person-years at risk**		**Incidence rate per 100,000 person-years** (95% CI)		
Age group (years)	Men	Women	Men	Women	All	Men	Women
1998	833	1044	7.19	7.58	127.05 (121.43 – 132.93)	115.82 (108.22 – 123.96)	137.71 (129.61 – 146.32)
1999	975	1264	8.21	8.62	132.98 (127.52 – 138.60)	118.68 (111.46 – 126.37)	146.60 (138.73 – 154.91)
2000	1249	1472	10.04	10.49	132.53 (127.64 – 137.60)	124.40 (117.69 – 131.50)	140.30 (133.31 – 147.65)
2001	1564	1942	12.41	12.93	138.30 (133.80 – 142.96)	126.00 (119.91 – 132.40)	150.11 (143.58 – 155.64)
2002	1832	2171	14.01	14.55	140.15 (135.88 – 144.56)	130.73 (124.88 – 136.86)	149.23 (143.08 – 155.71)
2003	2051	2461	16.13	16.70	137.47 (133.51 – 141.54)	127.19 (121.80 – 132.81)	147.40 (141.69 – 153.34)
2004	2732	3510	17.90	18.45	171.72 (167.51 – 176.03)	152.63 (147.02 – 158.47)	190.23 (184.04 – 196.63)
2005	2598	3456	19.39	19.97	153.81 (149.98 – 157.73)	133.95 (128.90 – 139.21)	173.09 (167.42 – 178.96)
2006	2568	3127	19.94	20.50	140.84 (137.23 – 144.54)	128.79 (123.90 – 133.86)	152.56 (147.31 – 158.00)
2007	2415	2823	20.18	20.68	128.19 (124.77 – 131.71)	119.67 (114.99 – 124.54)	136.51 (131.56 – 141.64)
2008	2354	2773	20.46	20.94	123.85 (120.51 – 127.29)	115.07 (110.51 – 119.81)	132.43 (127.59 – 137.45)
2009	2493	2884	20.56	21.05	129.20 (125.79 – 132.70)	121.23 (116.56 – 126.08)	136.98 (132.07 – 142.07)
2010	2515	2793	20.39	20.88	128.62 (125.21 – 132.13)	123.36 (118.64 – 128.28)	133.75 (128.88 – 138.80)
2011	2430	2801	19.96	20.50	129.27 (125.81 – 132.82)	121.71 (116.97 – 126.65)	136.62 (131.65 – 141.77)
2012	2331	2445	19.81	20.37	118.86 (115.54 – 122.28)	117.65 (112.97 – 122.53)	120.04 (115.38 – 124.90)
2013	2215	2422	18.98	19.56	120.33 (116.91 – 123.84)	116.72 (111.96 – 121.68)	123.82 (118.99 – 128.85)
2014	1897	2021	17.54	18.05	110.09 (106.69 – 113.59)	108.18 (103.42 – 113.16)	111.94 (107.16 – 116.93)
2015	1394	1495	15.55	16.01	91.52 (88.24 – 94.92)	89.62 (85.04 – 94.46)	93.36 (88.74 – 98.21)
2016	1023	1020	13.09	13.46	76.96 (73.70 – 80.37)	78.18 (73.53 – 83.12)	75.78 (71.27 – 80.58)
2017	691	690	11.64	11.93	58.58 (55.58 – 61.76)	59.37 (55.10 – 63.97)	57.82 (53.66 – 62.30)

Appendix D.4.2 Stroke incidence presented by age group and sex (1998 – 2017)

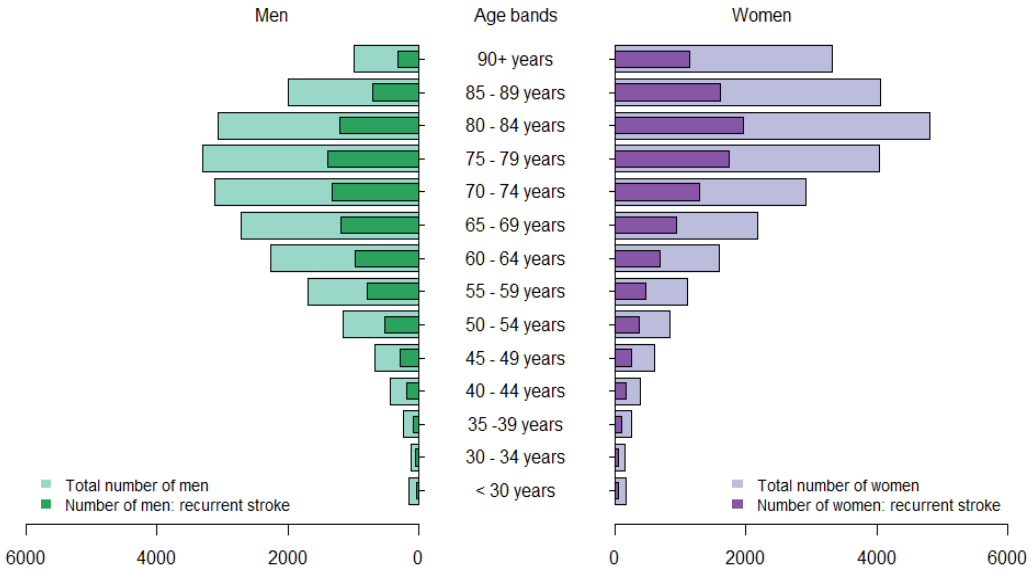
	Stroke events		**100,000 person-years at risk**		**Incidence rate per 100,000 person-years** (95% CI)		
Age group (years)	Men	Women	Men	Women	All	Men	Women
< 20	148	115	38.67	30.29	3.81 (3.38 – 4.30)	3.83 (3.26 – 4.50)	3.80 (3.16 – 4.56)
20 – 24	126	158	26.84	28.30	5.15 (4.58 – 5.79)	4.69 (3.94 – 5.59)	5.58 (4.78 – 6.52)
25 – 29	314	345	34.16	36.62	9.31 (8.63 – 10.05)	9.19 (8.23 – 10.27)	9.42 (8.48 – 10.47)
30 – 34	637	556	40.52	41.44	14.55 (13.75 – 15.40)	15.72 (14.54 – 16.99)	13.42 (12.35 – 14.58)
35 – 39	1126	858	41.22	40.13	24.39 (23.34 – 25.48)	27.31 (25.76 – 28.96)	21.38 (19.99 – 22.86)
40 – 44	1761	1155	38.77	36.70	38.64 (37.26 – 40.07)	45.42 (43.35 – 47.60)	31.48 (29.71 – 33.34)
45 – 49	2582	1629	34.07	32.68	63.09 (61.21 – 65.02)	75.79 (72.93 – 78.77)	49.84 (47.48 – 52.32)
50 – 54	3531	2345	30.27	29.83	97.78 (95.31 – 100.31)	116.66 (112.88 – 120.57)	78.62 (75.50 – 81.87)
55 – 59	4313	3299	25.96	26.67	144.63 (141.42 – 147.92)	166.12 (161.24 – 171.15)	123.71 (199.56 – 128.01)
60 – 64	4958	4453	20.90	22.96	214.56 (210.27 – 218.94)	237.21 (230.70 – 243.90)	193.94 (188.32 – 199.72)
65 – 69	5486	5864	16.85	19.78	309.88 (304.23 – 315.63)	325.63 (317.13 – 334.36)	296.47 (288.97 – 304.15)
70 – 74	5216	6784	11.86	15.75	434.71 (427.00 – 442.56)	439.87 (428.09 – 451.97)	430.82 (420.69 – 441.20)
75 – 79	4092	6807	7.25	11.47	582.02 (571.20 – 593.05)	564.04 (547.02 – 581.59)	593.39 (579.46 – 607.66)
80 – 84	2397	5262	3.59	7.15	713.37 (697.57 – 729.52)	668.01 (641.80 – 695.30)	736.14 (716.51 – 756.30)
85 – 89	1085	3182	1.37	3.59	860.57 (835.14 – 886.79)	790.09 (744.45 – 838.53)	887.57 (857.26 – 918.96)
90 – 94	329	1456	0.39	1.38	1000.00 (959.80 – 1100.00)	838.94 (753.01 – 934.67)	1100.00 (999.86 – 1100.00)
95+	59	346	0.11	0.41	783.57 (710.85 – 863.72)	548.94 (425.31 – 708.50)	845.16 (760.64 – 939.08)
All ages	38,160	44,614	372.80	385.14	109.21 (108.47 – 109.96)	102.36 (101.34 – 103.39)	115.84 (114.77 – 116.92)

Appendix D.4.3 **Distribution of subsequent major adverse outcomes by sex and 5-year age group for patients with subsequent outcome after 30 days of incident stroke (n=48,306)**

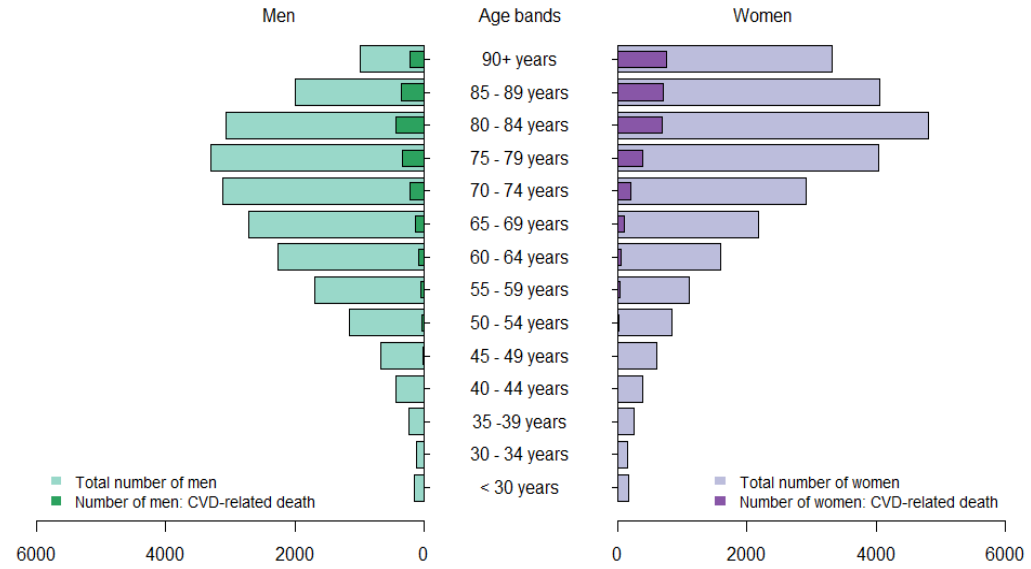
Major adverse cardiovascular disease (n=28,750)



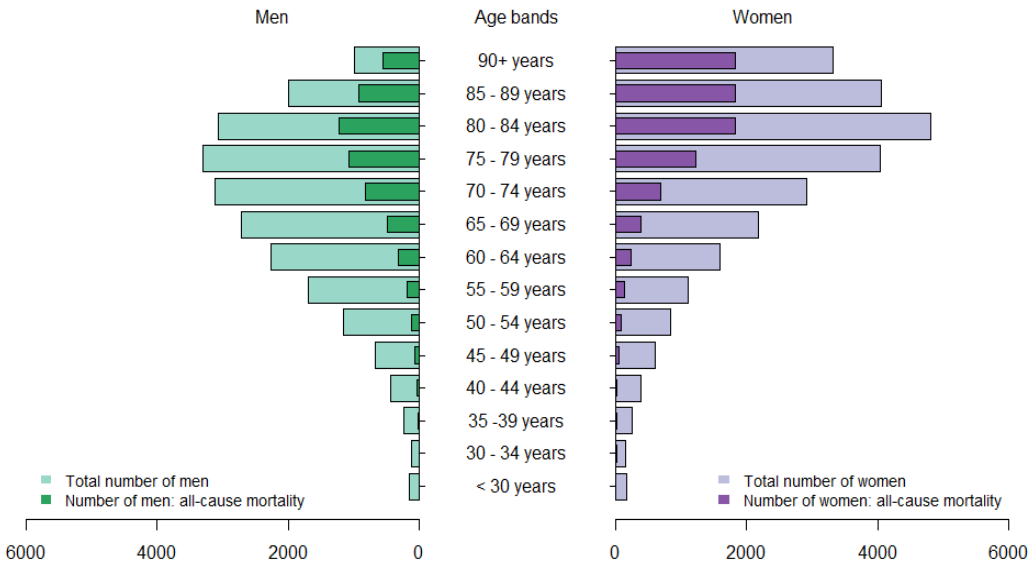
Recurrent stroke (n=19,896)



Cardiovascular-related mortality (n=4,797)



All-cause mortality (n=14,137)



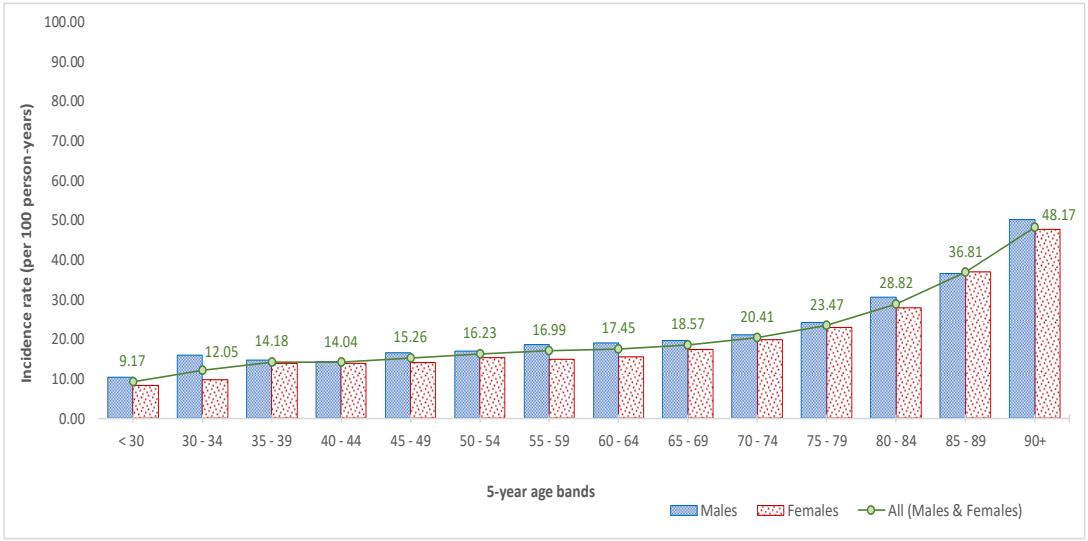
Appendix D.4.4 Descriptive characteristics of patients with subsequent outcome after 30 days of incident stroke by sex

	Total n = 48,306	**Men** n =21,906 (45.4%)	**Women** n =26,400 (54.7%)	***p*-value**
Age at incident stroke	73.1 (13.9)	70.3 (13.4)	75.5 (13.9)	0.0001
Age at subsequent MACE outcome	76.3 (13.1)	73.3 (12.8)	78.8 (12.9)	0.0001
Socioeconomic status				0.646
1 (Highest SES)	10,322 (21.4)	4,730 (21.6)	5,957 (22.6)	
2	10,789 (22.3)	4,832 (22.1)	5,957 (22.6)	
3	10,389 (21.5)	4,691 (21.4)	5,698 (21.6)	
4	8,878 (18.4)	4,018 (18.3)	4,860 (18.4)	
5 (Lowest SES)	7,847 (16.2)	3,597 (16.4)	4,250 (16.1)	
Missing	81 (0.2)	38 (0.2)	43 (0.2)	
Ethnicity				<0.001
Asian	615 (1.3)	327 (1.5)	288 (1.1)	
Black	379 (0.8)	183 (0.8)	196 (0.7)	
Mixed	73 (0.2)	38 (0.2)	35 (0.1)	
Other	336 (0.7)	163 (0.7)	173 (0.7)	
White	43,179 (89.4)	19,644 (89.7)	23,535 (89.2)	
Unknown	3,724 (7.7)	1,551 (7.1)	2,173 (8.2)	
Comorbid conditions				
Atrial fibrillation	4,212 (8.7)	1,839 (8.4)	2,373 (9.0)	0.021
Diabetes mellitus	5, 495 (11.4)	2,739 (12.5)	2,756 (10.4)	<0.001
Dyslipidaemia	4,847 (10.0)	2,083 (9.5)	2,764 (10.5)	<0.001
Hypertension	22,461 (46.5)	9,402 (42.9)	13,059 (49.5)	<0.001
TIA	12,377 (25.6)	5,497 (25.1)	6,880 (26.1)	0.015
Outcomes				<0.001
Coronary heart disease	2,129 (4.4)	1,149 (5.2)	980 (3.7)	
Haemorrhagic stroke	930 (1.9)	463 (2.1)	467 (1.8)	
Ischaemic stroke	3,595 (7.4)	1,641 (7.5)	1,954 (7.4)	
Stroke (not specified)	15,371 (31.8)	6,944 (31.7)	8,427 (31.9)	
Peripheral vascular disease	533 (1.1)	301 (1.4)	232 (0.9)	
Heart failure	1,395 (2.9)	593 (2.7)	802 (3.0)	
CVD-related death	4,797 (9.9)	1,882 (8.6)	2,915 (11.0)	
Non-CVD related death	9,340 (19.3)	3,954 (18.0)	5,386 (20.4)	

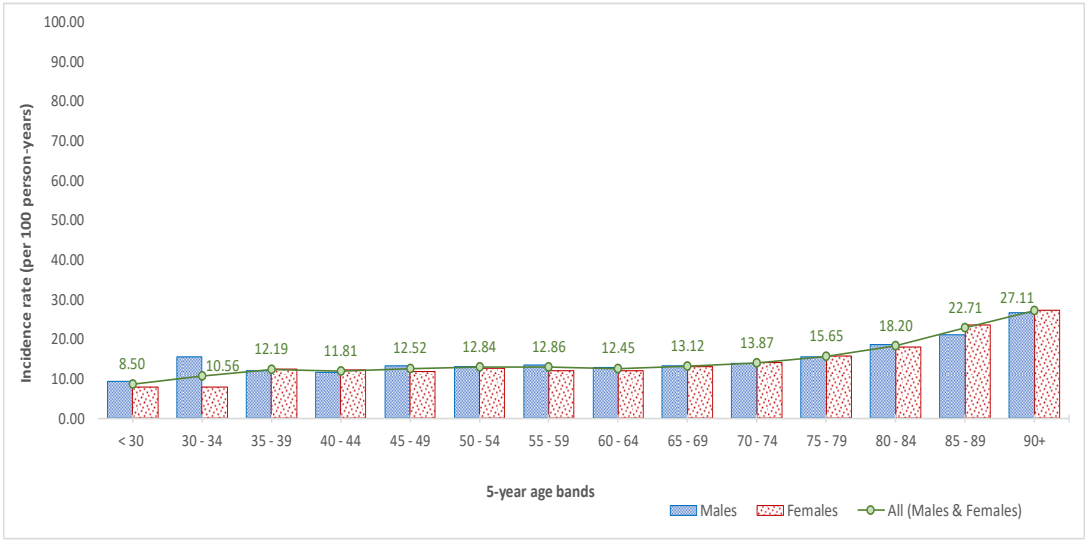
n: total number; %: percentage/proportion; Mean age for incident stroke and mean age for subsequent MACE outcome reported with standard deviation.

Appendix D.4.5 Incidence of subsequent major adverse outcomes presented by sex and 5-year age groups for patients with subsequent major adverse event after 30 days of index stroke (n=48,306)

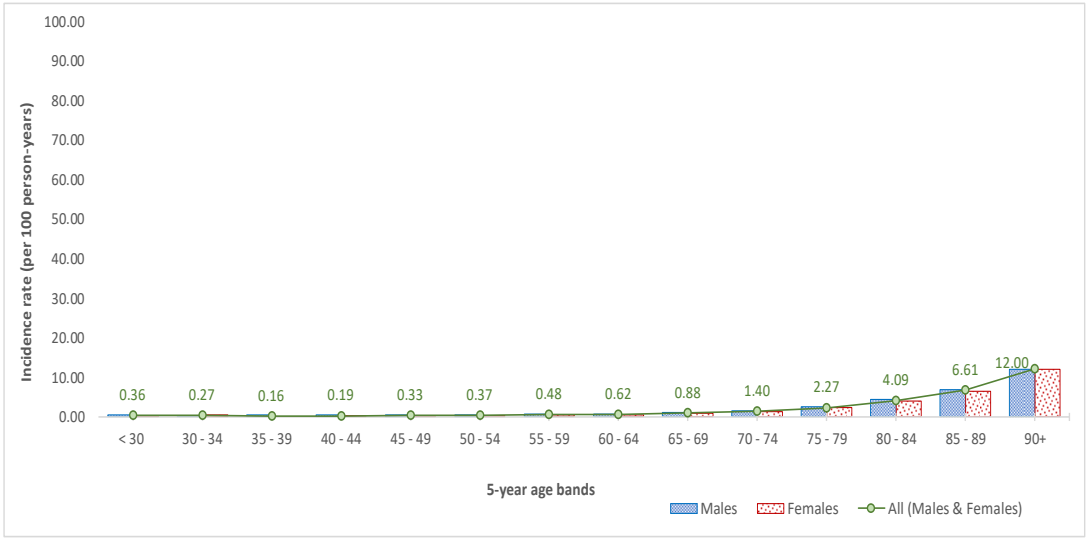
Major adverse cardiovascular events



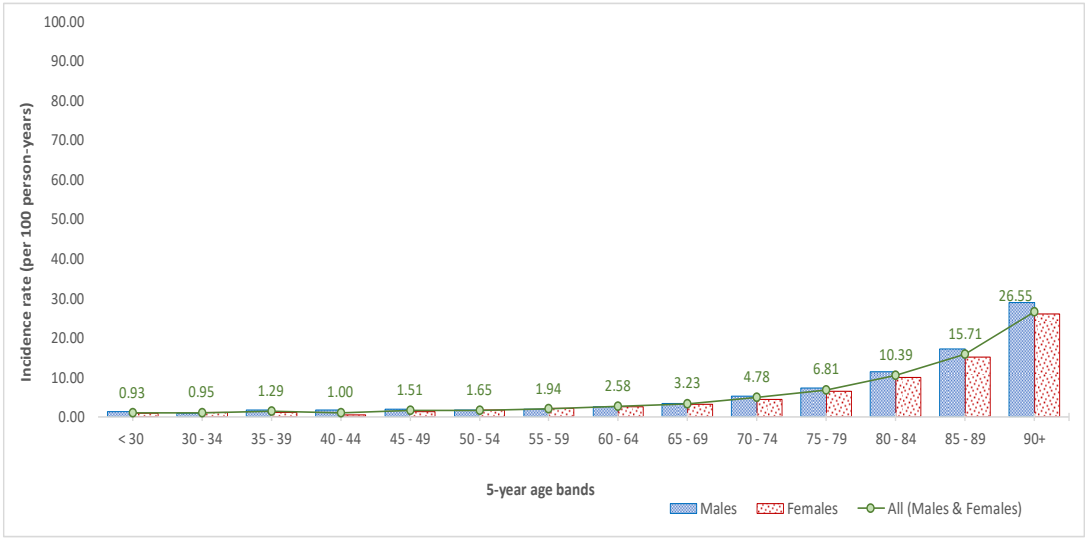
Recurrent stroke



Cardiovascular-related mortality



All-cause mortality



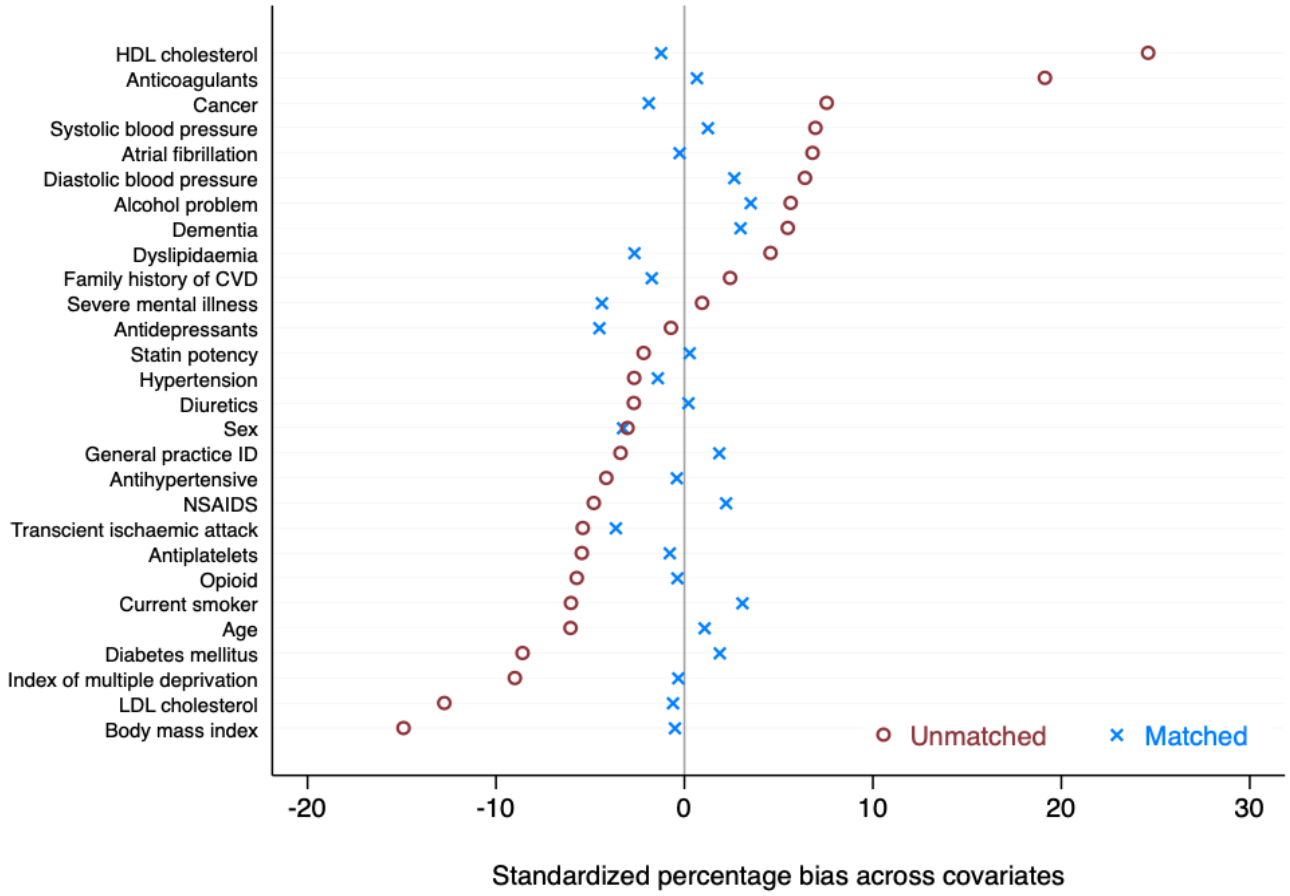
Appendix E Additional Results from Chapter 5

Comparison of risk of serious cardiovascular events after haemorrhagic versus ischaemic stroke

Appendix E.5.1 Number (proportion) of people with missing data on risk factors, by incident stroke sub-type

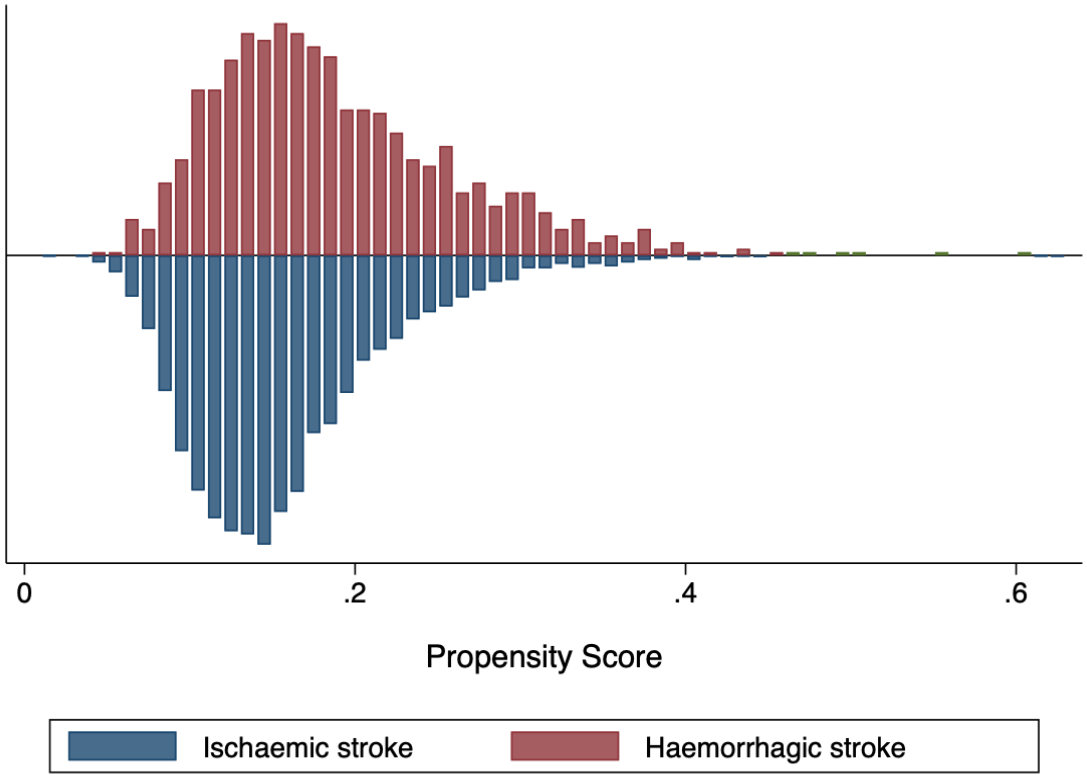
Variables	All n=32,091	Haemorrhagic [n=6,535 (20.4)]	Ischaemic [n=25,556 (79.6)]
Body mass index (kg/m^2)	17,729 (55.3)	3,907 (59.8)	13,822 (54.1)
High density lipoprotein cholesterol (mmol/L)	20,011 (62.4)	4,393 (67.2)	15,618 (61.1)
Low density lipoprotein cholesterol (mmol/L)	22,428 (69.9)	4,833 (74.0)	17,595 (68.9)
Total cholesterol (mmol/L)	16,989 (52.9)	3,853 (59.0)	13,136 (51.4)
Diastolic blood pressure (mmHg)	6,224 (19.4)	1,620 (24.8)	4,604 (18.0)
Systolic blood pressure (mmHg)	6,224 (19.4)	1,620 (24.8)	4,604 (18.0)

Appendix E.5.2 **The standardised percentage bias of the 28 selected baseline variables before and after propensity score matching for complete case cohort**

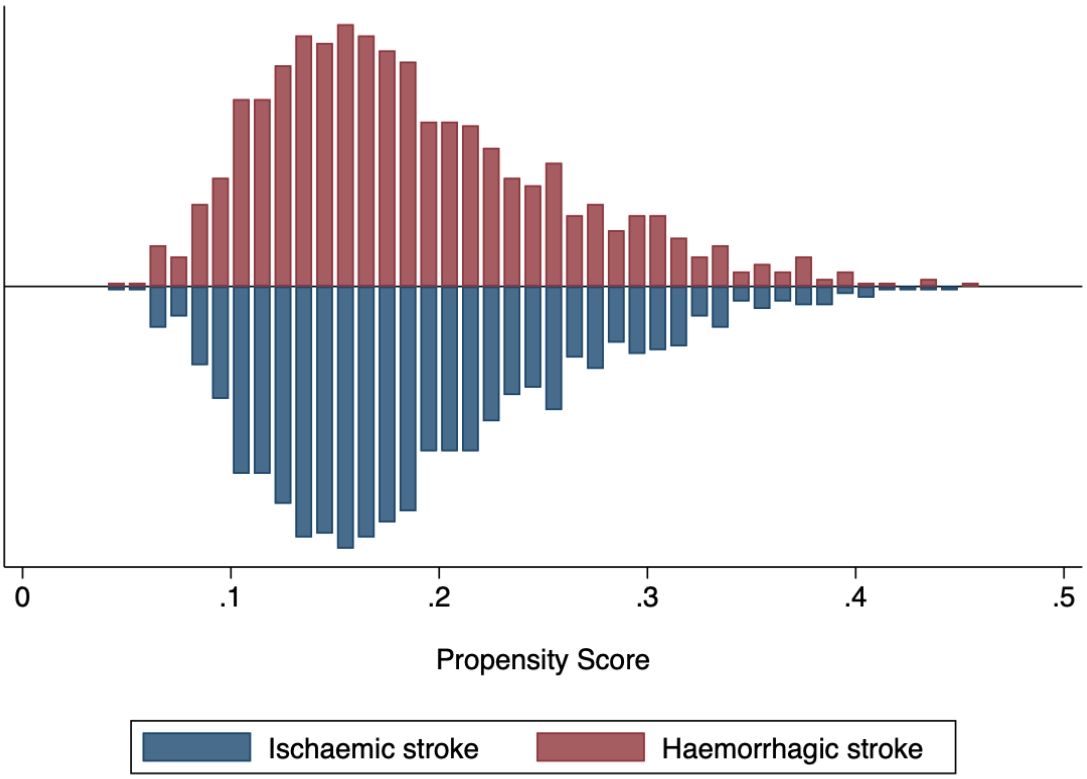


Appendix E.5.3 Distribution of propensity score before and after matching for complete case cohort

(a) Distribution of propensity score before matching



(b) Distribution of propensity score after matching

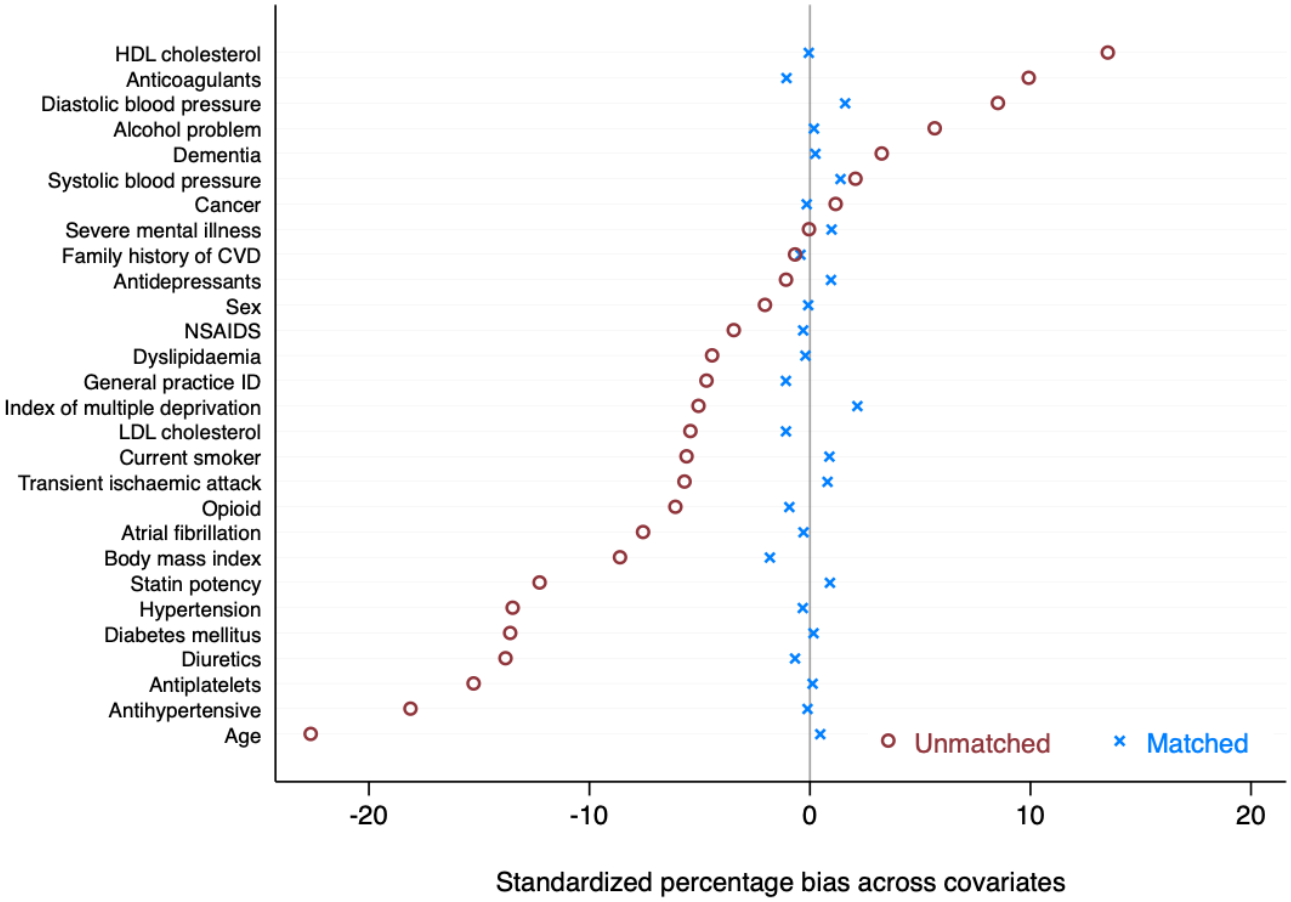


Appendix E.5.4 Characteristics of propensity score-matched complete-case cohort at the time of incident stroke according to stroke subtype (n=6,413)

Outcomes	Ischaemic (n = 1,039)	Haemorrhagic (n= 1,039)	p-value
Females	525 (50.5)	508 (48.9)	0.456
Age (years), median (IQR)	74 (65 – 81)	74 (66 – 81)	0.7693
Socioeconomic status			0.064
1 (Least deprived)	226 (21.8)	261 (25.1)	
2	229 (22.0)	209 (20.1)	
3	248 (23.87)	220 (21.2)	
4	195 (18.8)	178 (17.1)	
5 (Most deprived)	141 (13.6)	169 (16.3)	
Unknown	0	2 (0.2)	
Current smoker	155 (14.9)	167 (16.1)	0.467
Alcohol problem	39 (3.8)	46 (4.4)	0.438
Atrial fibrillation	162 (15.6)	161 (15.5)	0.952
Cancer	237 (22.8)	229 (22.0)	0.674
Dementia	41 (4.0)	47 (4.5)	0.513
Diabetes mellitus	305 (29.4)	314 (30.2)	0.666
Dyslipidaemia	222 (21.4)	211 (20.3)	0.552
Family history of CVD	283 (27.2)	275 (26.5)	0.692
Hypertension	695 (66.9)	688 (66.2)	0.745
Severe mental illness	25 (2.4)	19 (1.8)	0.361
Transient ischaemic attack	81 (7.8)	71 (6.8)	0.400
Anti-coagulant	129 (12.4)	131 (12.6)	0.895
Anti-depressant	265 (25.5)	245 (23.6)	0.308
Anti-hypertensive	727 (70.0)	725 (69.8)	0.924
Anti-platelet	379 (36.5)	375 (36.1)	0.855
Diuretics	421 (40.5)	422 (40.6)	0.964
NSAIDS	250 (24.1)	260 (25.0)	0.610
Opioids	433 (41.7)	431 (41.5)	0.929
Statin			0.515
Low intensity	63 (6.1)	66 (6.4)	
Moderate intensity	370 (35.6)	346 (33.3)	
High intensity	84 (8.1)	100 (9.6)	
Diastolic Blood Pressure (mmHg)	80 (71 – 85)	80 (71 – 85)	0.7125
Systolic Blood Pressure (mmHg)	140 (130 – 150)	140 (130 – 150)	0.9181
HDL cholesterol (mmol/L)	1.41 (1.19 – 1.80)	1.43 (1.20 – 1.80)	0.9154
LDL cholesterol (mmol/L)	2.70 (2.00 – 3.40)	2.70 (2.00 – 3.49)	0.9532
Total cholesterol (mmol/L)	4.90 (4.10 – 5.70)	4.80 (4.10 – 5.60)	0.4190

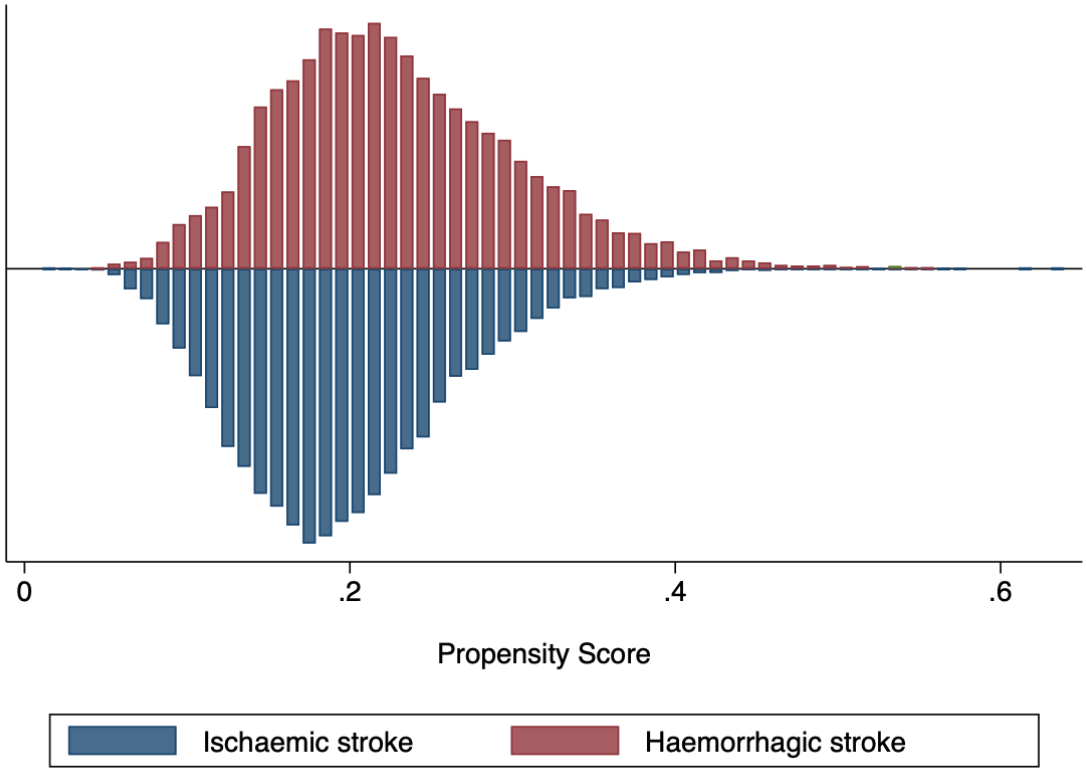
HDL: high density lipoprotein; LDL: low density lipoprotein; n: frequency/numbers; NSAIDS: non-steroidal anti-inflammatory drug; %: percent

Appendix E.5.5 **The standardised percentage bias of the 28 selected baseline variables before and after propensity score matching for the entire cohort**

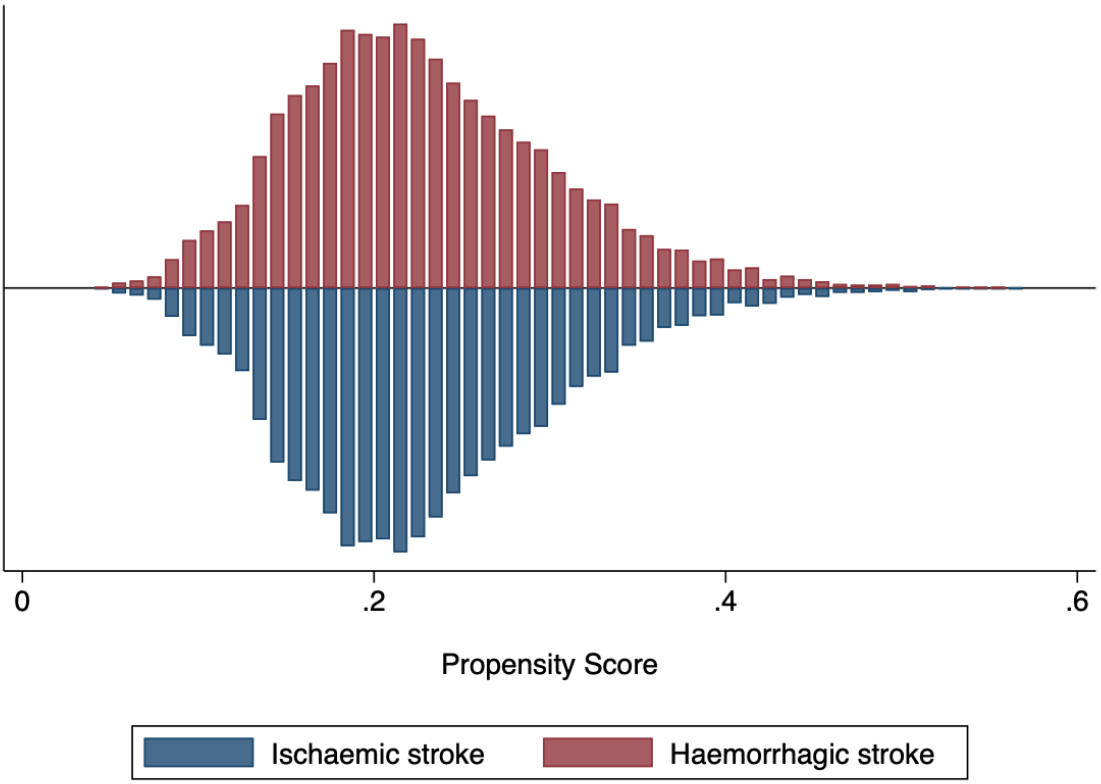


Appendix E.5.6 Distribution of propensity score before and after matching for the entire cohort

(a) Distribution of propensity score before matching



(b) Distribution of propensity score after matching



Appendix E.5.7 Characteristics of propensity score-matched cohort at the time of incident stroke according to stroke subtype (entire cohort, n=32,091)

Outcomes	Ischaemic (n = 6,534)	Haemorrhagic (n= 6,534)	p-value
Females	3,377 (51.7)	3,374 (51.6)	0.958
Age (years), median (IQR)	72 (60 – 82)	73 (61 – 82)	0.1987
Socioeconomic status			0.288
1 (Least deprived)	1,581 (24.2)	1,513 (23.2)	
2	1,438 (22.0)	1,450 (22.2)	
3	1,375 (21.0)	1,388 (21.2)	
4	1,157 (17.7)	1,183 (18.1)	
5 (Most deprived)	978 (15.0)	986 (15.1)	
Unknown	5 (0.1)	14 (0.2)	
Current smoker	1,110 (17.0)	1,132 (17.3)	0.610
Alcohol problem	259 (4.0)	261 (4.0)	0.929
Atrial fibrillation	591 (9.0)	585 (9.0)	0.854
Cancer	1,117 (17.1)	1,113 (17.0)	0.926
Dementia	284 (4.4)	287 (4.4)	0.898
Diabetes mellitus	573 (8.8)	576 (8.8)	0.926
Dyslipidaemia	518 (7.9)	514 (7.9)	0.897
Family history of CVD	1,150 (17.6)	1,139 (17.4)	0.800
Hypertension	2,732 (41.8)	2,721 (41.6)	0.845
Severe mental illness	74 (1.1)	81 (1.2)	0.572
Transient ischaemic attack	327 (5.0)	339 (5.2)	0.633
Anti-coagulant	561 (8.6)	543 (8.3)	0.571
Anti-depressant	1,327 (20.3)	1,352 (20.7)	0.588
Anti-hypertensive	2,733 (41.8)	2,729 (41.8)	0.943
Anti-platelet	1,445 (22.1)	1,448 (22.2)	0.950
Diuretics	1,903 (29.1)	1,882 (28.8)	0.685
NSAIDS	1,555 (23.8)	1,546 (23.7)	0.853
Opioids	2,394 (36.6)	2,364 (36.2)	0.585
Statin			0.908
Low intensity	160 (2.5)	165 (2.5)	
Moderate intensity	796 (12.2)	799 (12.2)	
High intensity	192 (2.9)	205 (3.1)	
Diastolic Blood Pressure (mmHg)	80 (76 – 84)	80 (76 – 84)	0.8098
Systolic Blood Pressure (mmHg)	140 (131 – 148)	140 (132 – 148)	0.3576
HDL cholesterol (mmol/L)	1.49 (1.34 – 1.66)	1.49 (1.35 – 1.65)	1.000
LDL cholesterol (mmol/L)	2.96 (2.70 – 3.21)	2.97 (2.70 – 3.21)	0.5231
Total cholesterol (mmol/L)	5.10 (4.79 – 5.40)	5.10 (4.80 – 5.40)	0.7702

HDL: high density lipoprotein; LDL: low density lipoprotein; n: frequency/numbers; NSAIDS: non-steroidal anti-inflammatory drug; %: percent

Appendix F Additional Results from Chapter 6

Obesity and long-term outcomes after incident stroke: a prospective population-based cohort study

Appendix F.6.1 Number (proportion) of people with missing data on risk factors, by sex

Variables	All n=30,702	Men n=14,303 (46.6%)	Women n=16,399 (53.4%)
Low density lipoprotein cholesterol (mmol/L)	16,135 (52.6)	7,176 (50.2)	8,959 (54.6)
High density lipoprotein cholesterol (mmol/L)	12,816 (41.7)	5,596 (39.1)	7,220 (44.0)
Total cholesterol (mmol/L)	8,989 (29.3)	3,760 (26.3)	5,229 (31.9)
Diastolic blood pressure (mmHg)	963 (3.1)	475 (3.3)	488 (3.0)
Systolic blood pressure (mmHg)	963 (3.1)	475 (3.3)	488 (3.0)
Glomerular filtration rate	17,613 (57.4)	8,223 (57.5)	9,390 (57.3)

Appendix F.6.2 Descriptive characteristics comparing patients with or within BMI record within 24 months of incident stroke

Characteristics	Patient with BMI n (%) - 30,702 (44.7)	Patient without BMI n (%) - 37,940 (55.3)	*p*-value
Age (years), Median (IQR)	75 (65 – 82)	77 (65 – 85)	0.0001
Age (years), Mean (SD)	72.5 (13.1)	74.0 (14.6)	
Female	16,399 (53.4)	20,854 (55.0)	<0.001
Comorbidities and risk factors			
Diastolic blood pressure, mmHg	80 (70 – 85)	80 (76 – 83)	0.0001
Systolic blood pressure, mmHg	140 (130 – 150)	140 (134 – 148)	0.0001
Total cholesterol, mmol/L	5.0 (4.4 – 5.6)	5.1 (4.9 – 5.4)	0.0001
Alcohol problem, mmol/L	967 (3.2)	936 (2.5)	<0.001
Atrial fibrillation	3,309 (10.8)	3,144 (8.3)	<0.001
Chronic kidney disease	4,462 (14.5)	2,770 (7.3)	<0.001
Diabetes mellitus	6,852 (22.3)	1,126 (3.0)	<0.001
Dyslipidaemia	4,121 (13.4)	2,439 (6.4)	<0.001
Hypertension	17,511 (57.0)	14,333 (37.8)	<0.001
Current smoker	5,689 (18.5)	6,413 (16.9)	<0.001
Transient ischaemic attack	6,998 (22.8)	7,070 (18.6)	<0.001
Medication prescriptions			
ACE inhibitor	12,384 (40.0)	7,861 (20.7)	<0.001
Anti-hypertensive	18,433 (60.0)	14,914 (39.3)	<0.001
Anti-diabetic	5,506 (17.9)	862 (2.3)	<0.001
Anti-platelet	13,247 (43.2)	12,429 (32.8)	<0.001
Beta-blocker	8,151 (26.6)	7,542 (19.9)	<0.001
Calcium channel blocker	9,332 (30.4)	7,161 (18.9)	<0.001
NSAIDS	8,329 (27.1)	9,250 (24.4)	<0.001
Statin			<0.001
Low intensity	1,536 (5.0)	904 (2.4)	
Moderate intensity	8,129 (26.5)	4,382 (11.6)	
High intensity	2,074 (6.8)	843 (2.2)	

ACE: angiotensin-converting enzyme; IQR: interquartile range; SD: standard deviation.

Appendix F.6.3 Descriptive characteristics of patients with body mass index record before incident stroke stratified by sex

Characteristics	Men n (%) 14,303 (46.6)	Women n (%) 16,399 (53.4)	*p*-value
Age (years), Median (IQR)	72 (63 – 80)	76 (67 – 84)	0.0001
Age (years), Mean (SD)	71.0 (12.3)	73.8 (13.6)	
Comorbidities and risk factors			
Diastolic blood pressure, mmHg	80 (71 – 85)	80 (70 – 84)	0.0001
Systolic blood pressure, mmHg	140 (130 – 148)	140 (130 – 150)	0.0151
Total cholesterol, mmol/L	4.8 (4.2 – 5.3)	5.2 (4.6 – 5.7)	0.0001
Alcohol problem	676 (4.7)	291 (1.8)	<0.001
Atrial fibrillation	1,502 (10.5)	1,807 (11.0)	0.145
Chronic kidney disease	1,820 (12.7)	2,642 (16.1)	<0.001
Diabetes mellitus	3,498 (24.5)	3,354 (20.5)	<0.001
Dyslipidaemia	1,784 (12.5)	2,337 (14.3)	<0.001
Hypertension	7,794 (54.5)	9,717 (59.3)	<0.001
Current smoker	3,009 (21.0)	2,680 (16.3)	<0.001
Transient ischaemic attack	3,330 (23.3)	3,668 (22.4)	0.057
Medication prescriptions			
ACE inhibitor	5,826 (40.7)	6,458 (39.4)	0.016
Anti-hypertensive	8,355 (58.4)	10,078 (61.5)	<0.001
Anti-diabetic	2,883 (20.2)	2,623 (16.0)	<0.001
Anti-platelet	6,196 (43.3)	7,051 (43.0)	0.569
Beta-blocker	3,373 (23.6)	4,778 (29.1)	<0.001
Calcium channel blocker	4,317 (30.2)	5,015 (30.6)	0.449
NSAIDS	3,685 (25.8)	4,644 (28.3)	<0.001
Statin			<0.001
Low intensity	724 (5.1)	812 (5.0)	
Moderate intensity	4,041 (28.3)	4,088 (24.9)	
High intensity	1,002 (7.0)	1,072 (6.5)	

ACE: angiotensin-converting enzyme; IQR: interquartile range; SD: standard deviation

Appendix F.6.4 Outcomes in body mass index subgroups presented by sex for the fully adjusted model

Outcomes		**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m^2** n=768 (2.5%)
		HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Composite MACE						
	Men	1.04 (0.90 – 1.21)	0.97 (0.92 – 1.02)	0.99 (0.93 – 1.05)	1.06 (0.96 – 1.17)	1.15 (0.98 – 1.34)
	Women	1.15 (1.05 – 1.25)	0.96 (0.92 – 1.00)	0.97 (0.92 – 1.03)	1.09 (1.01 – 1.19)	1.00 (0.89 – 1.12)
CHD						
	Men	0.48 (0.18 – 1.31)	0.98 (0.81 – 1.19)	1.10 (0.87 – 1.38)	1.11 (0.79 – 1.57)	0.60 (0.29 – 1.22)
	Women	0.97 (0.61 – 1.55)	0.91 (0.74 – 1.11)	1.00 (0.78 – 1.28)	1.12 (0.72 – 1.44)	1.00 (0.62 – 1.63)
Recurrent stroke						
	Men	0.94 (0.77 – 1.13)	1.02 (0.97 – 1.08)	1.03 (0.96 – 1.10)	1.05 (0.93 – 1.17)	1.23 (1.03 – 1.46)
	Women	1.02 (0.91 – 1.13)	1.01 (0.96 – 1.06)	1.01 (0.95 – 1.08)	1.13 (1.03 – 1.25)	1.00 (0.87 – 1.14)
PVD						
	Men	2.03 (0.87 – 4.75)	0.66 (0.45 – 0.95)	0.83 (0.53 – 1.28)	1.03 (0.54 – 1.94)	-
	Women	1.79 (0.93 – 3.45)	0.66 (0.44 – 1.00)	0.80 (0.49 – 1.30)	0.44 (0.19 – 1.05)	0.34 (0.08 – 1.43)
Heart failure						
	Men	1.23 (0.50 – 3.04)	0.87 (0.65 – 1.15)	1.36 (0.97 – 1.89)	2.13 (1.35 – 3.37)	1.60 (0.68 – 3.75)
	Women	1.13 (0.66 – 1.94)	1.24 (0.96 – 1.60)	1.44 (1.06 – 1.97)	2.09 (1.41 – 3.11)	2.29 (1.29 – 4.07)
Cardiovascular mortality						
	Men	1.58 (1.19 – 2.11)	0.81 (0.72 – 0.92)	0.80 (0.68 – 0.95)	0.74 (0.55 – 1.00)	1.23 (0.80 – 1.88)
	Women	1.51 (1.29 – 1.76)	0.79 (0.71 – 0.88)	0.78 (0.68 – 0.90)	0.84 (0.67 – 1.04)	0.92 (0.66 – 1.28)

Outcomes		**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m^2** n=768 (2.5%)
All-cause mortality						
	Men	1.69 (1.41 – 2.02)	0.73 (0.68 – 0.79)	0.73 (0.65 – 0.81)	0.80 (0.66 – 0.96)	0.90 (0.67 – 1.22)
	Women	1.64 (1.48 – 1.82)	0.77 (0.71 – 0.83)	0.77 (0.70 – 0.85)	0.74 (0.64 – 0.87)	1.04 (0.84 – 1.28)

Total number of men, 14,303 (46.6%); women, 16,399 (53.4%)

Model adjusted for age, socioeconomic status, current smoking, history of an alcohol problem, atrial fibrillation, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, prescription of ACE inhibitor, anti-hypertensive, anti-diabetic, anti-platelet, beta-blocker, calcium channel blocker, NSAIDS, statin potency, diastolic and systolic blood pressure, glomerular filtration rate, total cholesterol.

Reference category: Normal weight patients with a BMI of 18.5-24.9 kg/m^2.

Appendix F.6.5 **Outcomes in body mass index subgroups presented by diabetes mellitus status for the fully adjusted model**

Outcomes	**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m²** n=768 (2.5%)
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Composite MACE					
Non-diabetics	1.15 (1.06 – 1.24)	0.97 (0.94 – 1.01)	0.98 (0.93 – 1.03)	1.08 (1.00 – 1.17)	0.97 (0.86 – 1.09)
Diabetics	0.94 (0.74 – 1.20)	0.92 (0.85 – 0.99)	0.96 (0.88 – 1.05)	1.05 (0.94 – 1.18)	1.15 (0.98 – 1.35)
CHD					
Non-diabetics	0.86 (0.56 – 1.33)	0.93 (0.79 – 1.09)	0.99 (0.81 – 1.21)	1.13 (0.83 – 1.54)	0.99 (0.60 – 1.63)
Diabetics	0.66 (0.16 – 2.69)	1.01 (0.75 – 1.36)	1.17 (0.85 – 1.61)	0.95 (0.63 – 1.46)	0.63 (0.32 – 1.25)
Recurrent stroke					
Non-diabetics	1.01 (0.92 – 1.11)	1.03 (0.99 – 1.08)	1.03 (0.97 – 1.09)	1.09 (1.00 – 1.19)	0.99 (0.86 – 1.12)
Diabetics	0.89 (0.66 – 1.21)	0.93 (0.84 – 1.02)	0.95 (0.86 – 1.06)	1.04 (0.91 – 1.20)	1.17 (0.97 – 1.40)
PVD					
Non-diabetics	1.88 (1.06 – 3.34)	0.58 (0.41 – 0.82)	0.71 (0.46 – 1.12)	0.81 (0.40 – 1.64)	0.24 (0.03 – 1.73)
Diabetics	1.87 (0.56 – 6.23)	0.75 (0.47 – 1.19)	0.90 (0.54 – 1.50)	0.65 (0.31 – 1.37)	0.16 (0.02 – 1.20)
Heart failure					
Non-diabetics	1.04 (0.63 – 1.71)	1.03 (0.83 – 1.27)	1.42 (1.08 – 1.85)	1.81 (1.21 – 2.72)	2.99 (1.70 – 5.23)
Diabetics	1.72 (0.52 – 5.62)	1.16 (0.77 – 1.73)	1.45 (0.93 – 2.27)	2.54 (1.55 – 4.18)	1.12 (0.46 – 2.74)
Cardiovascular mortality					
Non-diabetics	1.59 (1.38 – 1.83)	0.77 (0.70 – 0.84)	0.74 (0.65 – 0.85)	0.82 (0.65 – 1.03)	0.96 (0.67 – 1.38)
Diabetics	0.98 (0.61 – 1.58)	0.89 (0.76 – 1.05)	0.91 (0.75 – 1.11)	0.80 (0.60 – 1.06)	1.15 (0.78 – 1.69)

Outcomes	**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m^2** n=768 (2.5%)
All-cause mortality					
Non-diabetics	1.66 (1.51 – 1.82)	0.73 (0.68 – 0.77)	0.74 (0.68 – 0.81)	0.78 (0.67 – 0.91)	0.99 (0.78 – 1.25)
Diabetics	1.50 (1.15 – 1.96)	0.83 (0.75 – 0.93)	0.79 (0.69 – 0.90)	0.76 (0.63 – 0.92)	1.04 (0.80 – 1.35)

The total number with a history of diabetes, 6,852 (22.3%); with no prior history of diabetes, 23,850 (77.7%).

Model adjusted for age, sex, socioeconomic status, current smoking, history of an alcohol problem, atrial fibrillation, chronic kidney disease, dyslipidaemia, hypertension, transient ischaemic attack, prescription of ACE inhibitor, anti-hypertensive, anti-diabetic, anti-platelet, beta-blocker, calcium channel blocker, NSAIDS, statin potency, diastolic and systolic blood pressure, glomerular filtration rate, total cholesterol.

Reference category: Normal weight patients with a BMI of 18.5-24.9 kg/m^2.

Appendix F.6.6 Outcomes in body mass index subgroups presented by smoking status for the fully adjusted model

Outcomes	**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m²** n=768 (2.5%)
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Composite MACE					
Non-current smoker	1.12 (1.03 – 1.23)	0.96 (0.93 – 1.00)	0.99 (0.95 – 1.04)	1.08 (1.00 – 1.15)	1.02 (0.92 – 1.14)
Current smoker	1.13 (0.98 – 1.30)	0.96 (0.89 – 1.04)	0.93 (0.84 – 1.03)	1.10 (0.95 – 1.27)	1.13 (0.92 – 1.40)
CHD					
Non-current smoker	0.81 (0.48 – 1.37)	0.93 (0.80 – 1.09)	1.06 (0.88 – 1.28)	1.09 (0.83 – 1.43)	1.00 (0.66 – 1.51)
Current smoker	0.88 (0.44 – 1.75)	1.02 (0.75 – 1.39)	1.08 (0.74 – 1.57)	1.00 (0.57 – 1.78)	0.17 (0.02 – 1.25)
Recurrent stroke					
Non-current smoker	0.98 (0.87 – 1.09)	1.01 (0.97 – 1.06)	1.04 (0.98 – 1.10)	1.08 (0.99 – 1.18)	1.04 (0.92 – 1.17)
Current smoker	1.05 (0.89 – 1.25)	1.03 (0.94 – 1.12)	0.92 (0.82 – 1.04)	1.13 (0.96 – 1.34)	1.17 (0.93 – 1.49)
PVD					
Non-current smoker	2.25 (1.19 – 4.28)	0.69 (0.49 – 0.96)	0.94 (0.64 – 1.38)	0.71 (0.38 – 1.32)	0.14 (0.02 – 0.99)
Current smoker	1.33 (0.56 – 3.16)	0.57 (0.35 – 0.94)	0.50 (0.26 – 0.98)	0.79 (0.32 – 1.93)	0.35 (0.05 – 2.60)
Heart failure					
Non-current smoker	1.09 (0.64 – 1.85)	1.11 (0.90 – 1.36)	1.37 (1.07 – 1.76)	2.15 (1.56 – 2.96)	1.93 (1.16 – 3.25)
Current smoker	1.16 (0.45 – 3.02)	0.68 (0.39 – 1.17)	1.69 (0.95 – 3.01)	1.94 (0.82 – 4.59)	2.41 (0.72 – 8.12)
Cardiovascular mortality					
Non-current smoker	1.52 (1.30 – 1.76)	0.80 (0.74 – 0.87)	0.77 (0.68 – 0.86)	0.81 (0.67 – 0.98)	0.93 (0.69 – 1.25)
Current smoker	1.51 (1.12 – 2.04)	0.76 (0.62 – 0.94)	0.91 (0.70 – 1.19)	0.80 (0.50 – 1.30)	1.50 (0.86 – 2.60)

Outcomes	**< 18.5** n=1,217 (4.0%)	**25.0 – 29.9** n=10,979 (35.8%)	**30.0 – 34.9** n=5,206 (17.0%)	**35.0 – 39.9** n=1,749 (5.7%)	**≥ 40 kg/m²** n=768 (2.5%)
All-cause mortality					
Non-current smoker	1.64 (1.48 – 1.82)	0.76 (0.72 – 0.81)	0.75 (0.69 – 0.81)	0.75 (0.66 – 0.86)	1.02 (0.84 – 1.23)
Current smoker	1.60 (1.33 – 1.93)	0.70 (0.62 – 0.80)	0.77 (0.65 – 0.92)	0.85 (0.64 – 1.14)	0.90 (0.59 – 1.37)

Total number of who are current smokers, 5,689 (18.5%); non-current smoker, 25,013 (81.5%).

Model adjusted for age, sex, socioeconomic status, diabetes mellitus, history of an alcohol problem, atrial fibrillation, chronic kidney disease, dyslipidaemia, hypertension, transient ischaemic attack, prescription of ACE inhibitor, anti-hypertensive, anti-diabetic, anti-platelet, beta-blocker, calcium channel blocker, NSAIDS, statin potency, diastolic and systolic blood pressure, glomerular filtration rate, total cholesterol.

Reference category: Normal weight patients with a BMI of 18.5-24.9 kg/m^2.

Appendix F.6.7 Outcomes in body mass index subgroups excluding patients with subsequent major adverse outcomes within 30 days of incident stroke (n=21,967)

Outcomes	**< 18.5** n=814 (3.7%)	**25.0 – 29.9** n=7,989 (36.4%)	**30.0 – 34.9** n=3,723 (17.0%)	**35.0 – 39.9** n=1,237 (5.6%)	**≥ 40 kg/m²** n=545 (2.5%)
Composite MACE					
Number of events (percent)	465 (57.1)	4,686 (58.7)	2,143 (57.6)	733 (59.3)	283 (51.9)
Follow-up time, years	0.97 (0.29 – 2.02)	1.14 (0.37 – 3.09)	1.18 (0.37 – 3.04)	1.20 (0.33 – 3.17)	1.08 (0.28 – 2.77)
Full adjustment HR (95% CI)	1.19 (1.08 – 1.31)	0.96 (0.92 – 1.00)	0.95 (0.90 – 1.00)	1.11 (1.02 – 1.20)	1.05 (0.93 – 1.18)
CHD					
Number of events (percent)	19 (2.3)	418 (5.2)	225 (6.0)	83 (6.7)	26 (4.8)
Follow-up time, years	1.28 (0.73 – 2.58)	2.24 (0.98 – 4.88)	2.09 (1.02 – 4.23)	3.14 (1.14 – 4.77)	1.64 (1.09 – 3.79)
Full adjustment HR (95% CI)	0.79 (0.50 – 1.26)	1.01 (0.87 – 1.16)	1.11 (0.93 – 1.32)	1.24 (0.97 – 1.60)	0.99 (0.66 – 1.50)
Recurrent stroke					
Number of events (percent)	293 (36.0)	3,375 (42.3)	1,492 (40.1)	494 (39.9)	204 (37.4)
Follow-up time, years	0.90 (0.27 – 1.72)	1.02 (0.31 – 2.38)	1.00 (0.28 – 2.33)	0.97 (0.23 – 2.07)	0.96 (0.22 – 2.33)
Full adjustment HR (95% CI)	1.10 (0.98 – 1.25)	1.00 (0.95 – 1.05)	0.95 (0.89 – 1.02)	1.03 (0.94 – 1.14)	1.01 (0.88 – 1.17)
PVD					
Number of events (percent)	17 (2.1)	86 (1.1)	59 (1.6)	19 (1.5)	2 (0.4)
Follow-up time, years	1.83 (1.07 – 2.76)	2.22 (0.90 – 5.02)	2.51 (0.81 – 4.84)	2.26 (1.22 – 3.74)	4.93 (1.23 – 8.63)
Full adjustment HR (95% CI)	1.07 (1.23 – 3.48)	0.63 (0.47 – 0.84)	0.86 (0.61 – 1.20)	0.83 (0.50 – 1.38)	0.24 (0.06 – 0.97)
Heart failure					
Number of events (percent)	19 (2.3)	233 (2.9)	124 (3.3)	58 (4.7)	20 (3.7)
Follow-up time, years	1.63 (0.86 – 3.43)	2.45 (0.89 – 4.90)	2.67 (0.88 – 5.96)	2.63 (0.83 – 5.05)	2.68 (1.30 – 5.84)
Full adjustment HR (95% CI)	1.21 (0.75 – 1.94)	1.12 (0.92 – 1.34)	1.42 (1.12 – 1.79)	2.31 (1.67 – 3.15)	2.36 (1.46 – 3.80)

Outcomes	**< 18.5** n=814 (3.7%)	**25.0 – 29.9** n=7,989 (36.4%)	**30.0 – 34.9** n=3,723 (17.0%)	**35.0 – 39.9** n=1,237 (5.6%)	**≥ 40 kg/m²** n=545 (2.5%)
Cardiovascular mortality					
Number of events (percent)	117 (14.4)	574 (7.2)	243 (6.5)	79 (6.4)	31 (5.7)
Follow-up time, years	0.98 (0.22 – 2.20)	1.91 (0.32 – 4.43)	2.04 (0.29 – 4.53)	2.14 (0.41 – 5.18)	0.90 (0.19 – 3.28)
Full adjustment HR (95% CI)	1.59 (1.31 – 1.94)	0.77 (0.69 – 0.86)	0.82 (0.71 – 0.96)	0.99 (0.78 – 1.26)	1.28 (0.89 – 1.85)
All-cause mortality					
Number of events (percent)	373 (45.8)	1,832 (22.9)	755 (20.3)	223 (18.0)	93 (17.1)
Follow-up time, years	1.40 (0.39 – 3.59)	2.62 (0.59 – 5.41)	2.67 (0.65 – 5.48)	2.21 (0.49 – 5.50)	1.24 (0.44 – 4.50)
Full adjustment HR (95% CI)	1.63 (1.46 – 1.82)	0.74 (0.70 – 0.79)	0.76 (0.70 – 0.83)	0.85 (0.74 – 0.98)	1.10 (0.89 – 1.35)

CHD: coronary heart disease; HR: hazards ratio; MACE: major adverse cardiovascular event; PVD: peripheral vascular disease.

Full adjustment for age, sex, socioeconomic status, current smoking, history of an alcohol problem, atrial fibrillation, chronic kidney disease, diabetes mellitus, dyslipidaemia, hypertension, transient ischaemic attack, prescription of ACE inhibitor, anti-hypertensive, anti-diabetic, anti-platelet, beta-blocker, calcium channel blocker, NSAIDS, statin potency, diastolic and systolic blood pressure, glomerular filtration rate, total cholesterol.

Reference category: Normal weight patients with a BMI of 18.5-24.9 kg/m^2.

Appendix G Additional Results from Chapter 7

A population-based study exploring phenotypic clusters and clinical outcomes in stroke using an unsupervised machine learning approach

Appendix G.7 Additional Methods

Feature selection

Least Absolute Shrinkage and Selection Operator (Lasso): Lasso is a linear regression-based model that is regularized by imposing an L1 penalty on the regression coefficients. The L1 penalty forces the sum of the absolute value of the coefficients to be less than a constant. The variable selection process is embedded in this model because, given the nature of the L1 norm, some coefficients will be forced to be 0, and hence are eliminated from the model.

Boruta: Boruta is a random forest-based method that iteratively removes the features that are proven to be statistically less relevant than random probes, which are artificial noise variables introduced in the model by the algorithm.

Kamila algorithm

The kamila algorithm[286] is a model-based adaptation of the k-means for managing heterogeneous (mixed) datasets. The Kamila algorithm begins with a set of centroids for the continuous variables and a set of parameters for the categorical variables. For continuous variables, the Euclidean distance with the closest centroid is computed. This set of N minimal distances is used to estimate the mixture distribution of continuous variables. For categorical variables, the probabilities of observing the data given the cluster are computed.

The log-likelihood of the sum of these two components is then used to find the most appropriate cluster for each subject. Based on this temporary partition, the centroids and the parameters are updated to best represent the clusters.

These steps are repeated until the clusters are stable. Finally, multiple runs of this process are performed with different initializations, and the partition maximizing the sum of the best final likelihoods is retained.

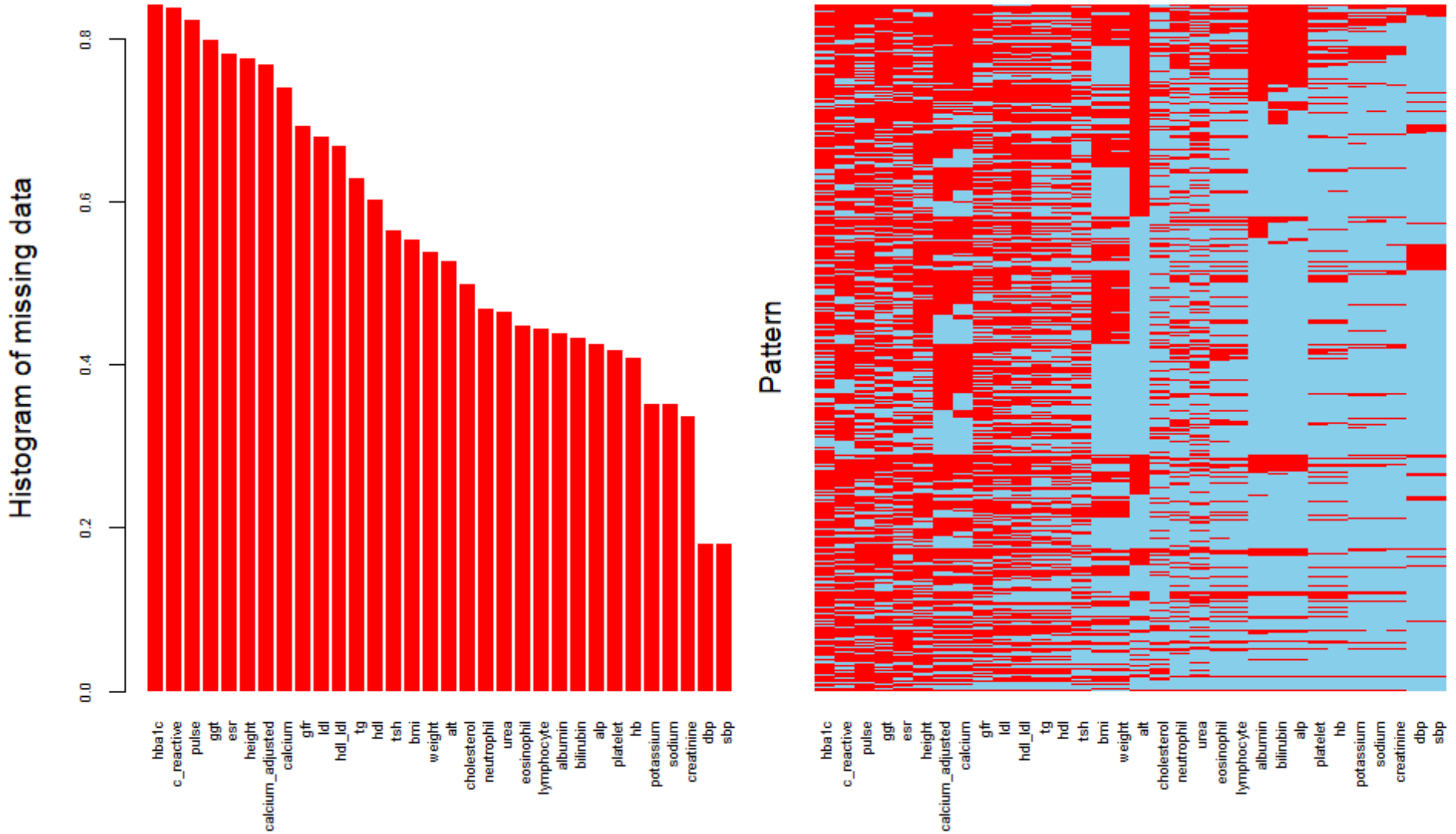
Gradient Boosting Model

The gradient boosting machines algorithm is a boosting algorithm that sequentially combines decision trees such that each additional tree is trained with more weighting placed on correctly predicting data points that the previous decision trees misclassified.[307] In simple terms, each new tree aims to correct the mistakes of the previous trees. Gradient boosting machines aim to minimise the loss function (a measure of the difference between the observed and predicted values) by combining a sequence of base-learner models. A common optimisation method to find a minimum is gradient descent which involves going down a gradient to reach a minimum. The key idea behind gradient boosting machines is to sequentially add a new base learner model to the ensemble sequence such that the new model is the model with the greatest correlation with the negative of the loss function's gradient calculated using the current ensemble sequence predictions.

SHAP (SHapley Additive exPlanations)

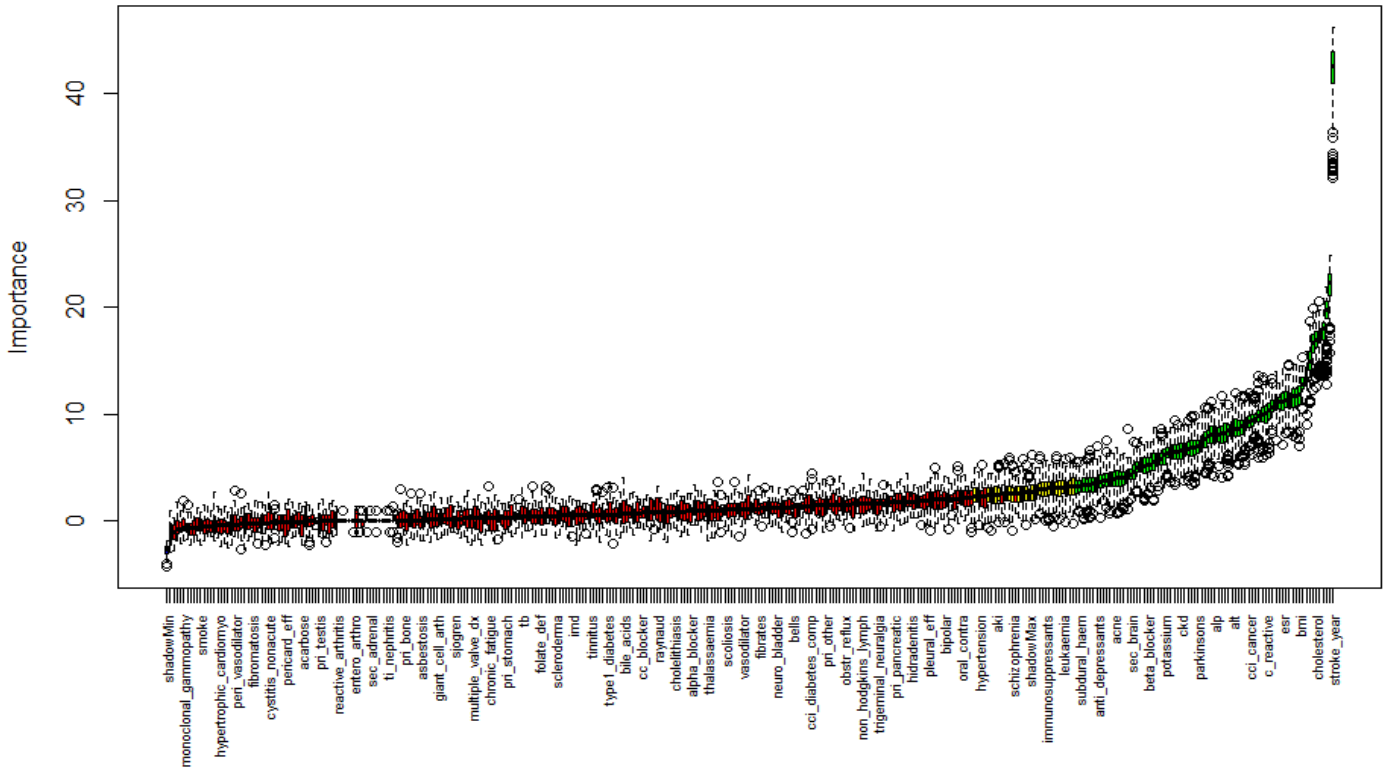
SHAP is a method to explain individual predictions and is based on the game theoretically optimal Shapley values. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The SHAP explanation method computes Shapley values from coalitional game theory. The feature values of a data instance act as players in a coalition. Shapley values indicate how to fairly distribute the "pay-out" (= the prediction) among the features/variables.

Appendix G.7 Figure 1 **All clinical variables with missing values**



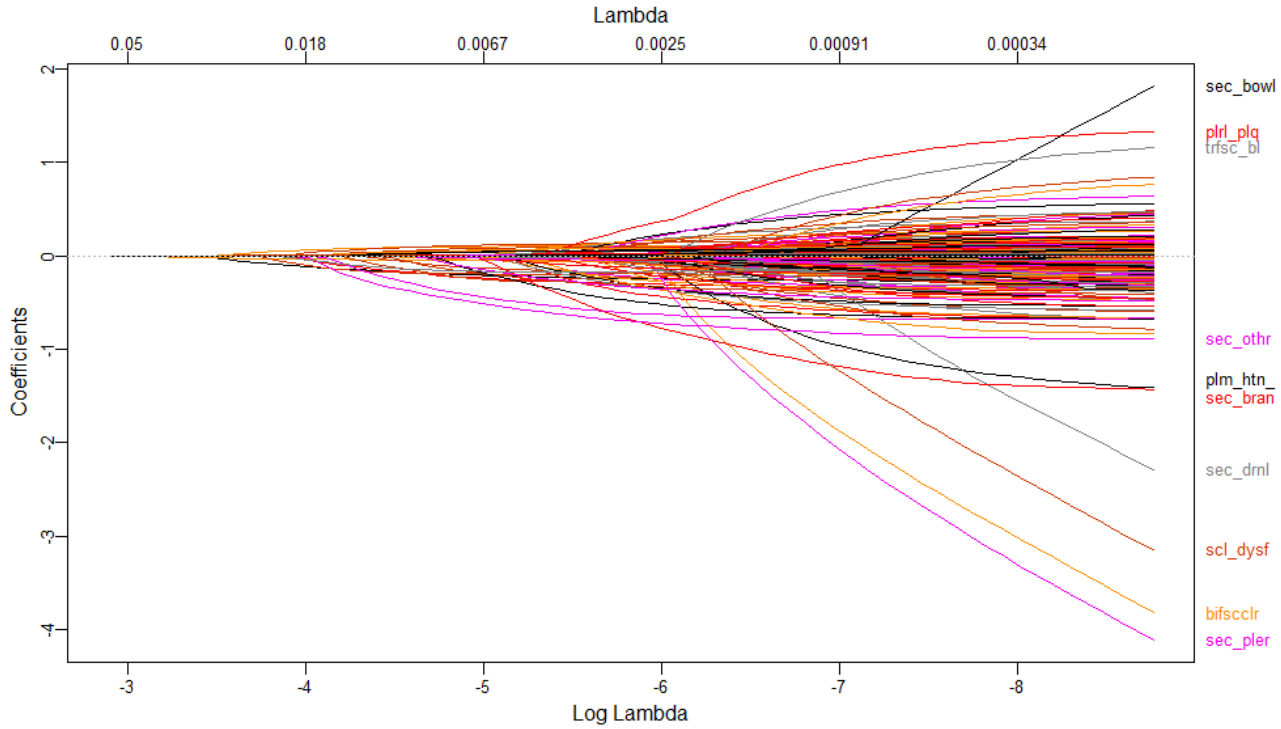
Appendix G.7 Figure 2 Feature selection

(a) Boruta – variable importance



This plot reveals the importance of each of the features. The columns in green are 'confirmed' and the ones in red are not.

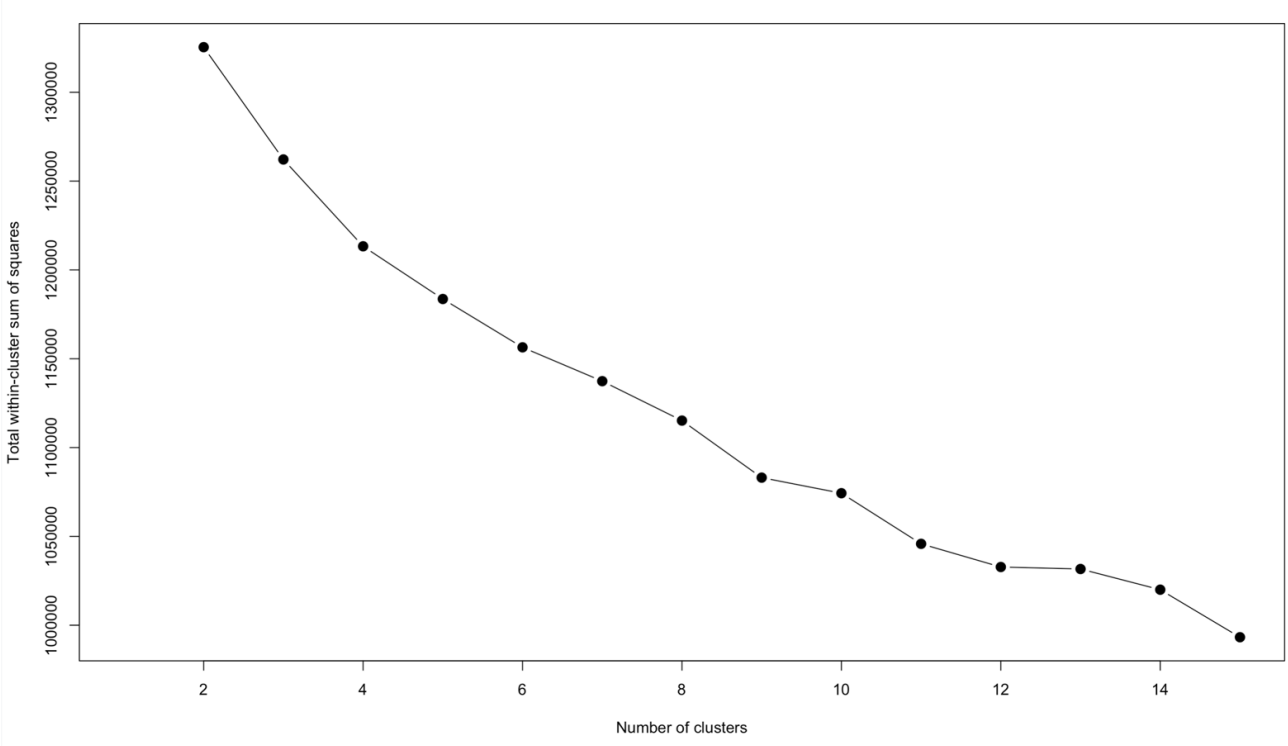
(b) LASSO regression



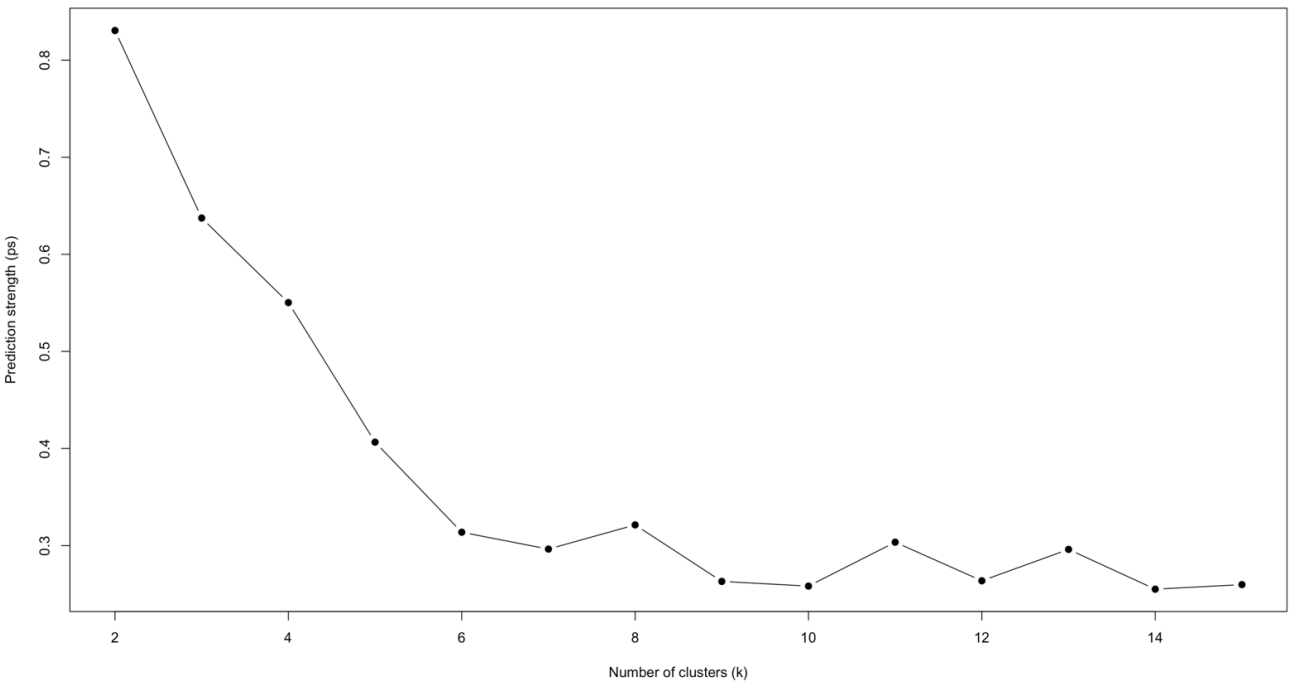
Each curve corresponds to a variable, showing the path of its coefficient against the ℓ1-norm of the whole coefficient vector as λ varies.

Appendix G.7 Figure 3 Optimal number of clusters

(a) Elbow method plot



(b) Prediction strength plot



Appendix G.7 Table 1 **Observed versus imputed values after multiple imputation for all clinical variables with missing data**

Variables	Median (Interquartile range)	
	Observed	Imputed
Alanine aminotransferase	19.0 (15.0 – 27.0)	23.2 (21.23 – 25.83)
Albumin level	41.0 (38.0 – 43.0)	40.6 (39.9 – 41.2)
Alkaline phosphatase	82.0 (67.0 – 104.0)	95.0 (89.1 – 102.5)
Bilirubin level	10.0 (7.0 – 13.0)	10.9 (10.1 – 11.8)
Body mass index	26.3 (23.1 – 30.0)	26.5 (25.6 – 27.4)
Diastolic blood pressure	80 (71 – 85)	80 (78 – 81)
Systolic blood pressure	140 (130 – 150)	141 (139 – 143)
Calcium level (adjusted)	2.34 (2.27 – 2.41)	2.34 (2.32 – 2.36)
Calcium level	2.33 (2.26 – 2.41)	2.34 (2.32 – 2.36)
Creatinine level	87.0 (74.0 – 104.0)	92.2 (88.6 – 97.0)
C-reactive protein	5.0 (3.0 – 11.0)	10.7 (7.3 – 15.3)
Eosinophil level	0.2 (0.2 – 0.3)	0.3 (0.2 – 0.4)
Erythrocyte sedimentation rate	14.0 (7.0 – 27.0)	18.5 (14.3 – 22.7)
Gamma glutamyl transpeptidase	29.0 (19.0 – 51.0)	44.9 (36 – 58.5)
Glomerular filtration rate	66.0 (56.0 – 81.0)	67.2 (64.0 – 70.5)
Haemoglobin level	13.5 (12.4 – 14.6)	13.5 (13.2 – 13.9)
Glycated haemoglobin (hba1c) level	47.5 (40.0 – 59.6)	50.1 (47.4 – 53.1)
HDL/LDL ratio	3.5 (2.0 – 4.4)	3.7 (3.4 – 4.0)
Height	1.65 (1.58 – 1.73)	1.67 (1.64 – 1.69)
High-density lipoprotein (HDL) cholesterol	1.4 (1.1 – 1.7)	1.5 (1.4 – 1.6)
Low-density lipoprotein (LDL) cholesterol	2.9 (2.2 – 3.6)	3.0 (2.8 – 3.2)
Lymphocyte count	1.8 (1.4 – 2.4)	3.2 (2.6 – 3.9)
Neutrophil count	4.3 (3.4 – 5.6)	4.9 (4.6 – 5.6)
Platelet count	248.0 (200.0 – 302.0)	248.0 (234.4 – 261.7)
Potassium level	4.4 (4.1 – 4.7)	4.4 (4.3 – 4.5)
Pulse	76 (68 – 84)	76 (74 – 79)
Sodium level	140 (137 – 142)	139 (138 – 140)
Thyroid-stimulating hormone level	1.8 (1.2 – 2.7)	2.1 (2.0 – 2.3)
Total cholesterol level	5.0 (4.2 – 5.8)	5.1 (4.9 – 5.3)
Triglyceride level	1.3 (1.0 – 1.8)	1.5 (1.3 – 1.6)
Urea	6.0 (4.8 – 7.6)	6.4 (5.9 – 6.9)
Weight	73.0 (61.6 – 85.0)	74.2 (70.8 – 77.6)

Appendix H Other publications during PhD studentship

1. Iyen, B., Vinogradova, Y., **Akyea, R.K.**, Weng, S., Qureshi, N., & Kai, J. (2022). Ethnic disparities in mortality among overweight or obese adults with newly diagnosed type 2 diabetes: a population-based cohort study. *Journal of Endocrinological Investigation*, https://doi.org/10.1007/s40618-021-01736-9

2. Weng, S., **Akyea, R.**, Man, K., Lau, W., Iyen, B., Blais, J., Chan, E., Siu, C., Qureshi, N., Wong, I. and Kai, J. (2021). Determining propensity for sub-optimal low-density lipoprotein cholesterol response to statins and future risk of cardiovascular disease. *PLOS ONE*, 16(12), https://doi.org/10.1371/journal.pone.0260839. *(Joint First Author)*

3. Qureshi, S., Latif, A., Condon, L., **Akyea, R.**, Kai, J. and Qureshi, N., (2021). Understanding the barriers and enablers of pharmacogenomic testing in primary care: a qualitative systematic review with meta-aggregation synthesis. *Pharmacogenomics*, https://doi.org/10.2217/pgs-2021-0131.

4. Iyen, B., **Akyea, R.K.**, Weng, S., Kai, J., & Qureshi, N. (2021). Statin treatment and LDL-cholesterol treatment goal attainment among individuals with familial hypercholesterolaemia in primary care. *Open Heart*, 8(2), https://doi.org/10.1136/openhrt-2021-001817

5. Qureshi, N., **Akyea, R.K.**, Dutton, B., Leonardi-Bee, J., Humphries, S. E., Weng, S., & Kai, J. (2021). Comparing the performance of the novel FAMCAT algorithms and established case-finding criteria for familial hypercholesterolaemia in primary care. *Open Heart*, 8(2), https://doi.org/10.1136/openhrt-2021-001752 *(Joint First Author)*

6. Qureshi, N., **Akyea, R.K.**, Dutton, B., Leonardi-Bee, J., Humphries, S. E., Weng, S., & Kai, J. (2021). Case-finding and Genetic testing for Familial Hypercholesterolaemia in Primary Care. *Heart*, https://doi.org/10.1136/heartjnl-2021-319742 *(Joint First Author)*

7. **Akyea, R.K.**, Kontopantelis, E., Kai, J., Weng, S. F., Patel, R. S., Asselbergs, F. W., & Qureshi, N. (2021). Sex disparity in subsequent outcomes in survivors of coronary heart disease. *Heart*, https://doi.org/10.1136/heartjnl-2021-319566

8. Iyen, B., Weng, S., Vinogradova, Y., **Akyea, R.K.**, Qureshi, N., & Kai, J. (2021). Long-term body mass index changes in overweight and obese adults and the risk of heart failure, cardiovascular disease and mortality: a cohort study of over 260,000 adults in the UK. *BMC Public Health*, 21, https://doi.org/10.1186/s12889-021-10606-1

9. Blais, J. E., **Akyea, R.K.**, Coetzee, A., Chan, A. H., Lau, W. C., Man, K. K., …Weng, S. (2021). Lipid levels and major adverse cardiovascular events in patients initiated on statins for primary prevention: an international population-based cohort study protocol. *BJGP Open*, 5(1), 1-9. https://doi.org/10.3399/bjgpopen20x101127 *(Joint First Author)*

10. **Akyea, R.K.,** Qureshi, N., Kai, J., de Lusignan, S., Sherlock, J., McGee, C., & Weng, S. (2020). Evaluating a clinical tool (FAMCAT) for identifying familial hypercholesterolaemia in primary care: a retrospective cohort study. *BJGP Open*, 4(5), https://doi.org/10.3399/bjgpopen20x101114

11. **Akyea, R.K.**, Qureshi, N., Kai, J., & Weng, S. F. (2020). Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care. *npj Digital Medicine*, 3(1), https://doi.org/10.1038/s41746-020-00349-5

12. Perkins, P., Parkinson, A., Parker, R., Blaken, A., & **Akyea, R.K.** (2020). Does acupressure help reduce nausea and vomiting in palliative care patients? A double-blind randomised controlled trial. *BMJ Supportive and Palliative Care*, https://doi.org/10.1136/bmjspcare-2020-002434

13. Ntaios, G., Weng, S., Perlepe, K., **Akyea, R.**, Condon, L., Lambrou, D., Sirimarco, G., Strambo, D., Eskandari, A., Karagkiozi, E., Vemmou, A., Korompoki, E., Manios, E., Makaritsis, K., Vemmos, K. and Michel, P., 2020. Data-driven machine-learning analysis of potential embolic sources in embolic stroke of undetermined source. *European Journal of Neurology*, https://doi.org/10.1111/ene.14524

14. Perkins P, Parkinson A, **Akyea RK**, Husbands E. Nasal fentanyl alone plus buccal midazolam: an open-label, randomised, controlled feasibility study in the dying. *BMJ Supportive and Palliative Care*, 10(3), 300-303. https://doi.org/10.1136/bmjspcare-2019-002029

15. **Akyea, R.K.**, Kai, J., Qureshi, N., Iyen, B. and Weng, S.F., 2019. LDL cholesterol response to statins and future risk of cardiovascular disease. *Heart*, 105(16), 1290-1291. https://doi.org/10.1136/heartjnl-2019-315461

16. Iyen, B., Qureshi, N., Kai, J., **Akyea, R.**, Leonardi-Bee, J., Roderick, P., Humphries, S.E. and Weng, S., 2019. Risk of cardiovascular disease outcomes in primary care subjects with familial hypercholesterolaemia: A cohort study. *Atherosclerosis*, 287, 8-15. https://doi.org/10.1016/j.atherosclerosis.2019.05.017

17. Weng, S., Kai, J., **Akyea, R.**, & Qureshi, N. (2019). Detection of familial hypercholesterolaemia: external validation of the FAMCAT clinical case-finding algorithm to identify patients in primary care. *Lancet Public Health*, 4(5), e256-e264. https://doi.org/10.1016/S2468-2667%2819%2930061-1

18. **Akyea, R.**, Kai, J., Qureshi, N., Iyen, B., & Weng, S. (2019). Sub-optimal cholesterol response to initiation of statins and future risk of cardiovascular disease. *Heart*, 105(13), 975-981. https://doi.org/10.1136/heartjnl-2018-314253

    (Awarded Heart Best Paper for 2019)

19. Patel, M., Lee, S.I., **Akyea, R.K.**, Grindlay, D., Francis, N., Levell, N.J., Smart, P., Kai, J., Thomas, K.S. (2019). A systematic review showing the lack of diagnostic criteria and tools developed for lower limb cellulitis. *British Journal of Dermatology*, 181(6), 1156-1165. https://doi.org/10.1111/bjd.17857