Evaluating genetic liability in hypertension and stroke using machine learning and traditional statistical models: Insights from UK biobank studies

A Thesis Submitted for the Degree of Doctor of Philosophy

By

Gideon MacCarthy

Department of Biosciences, Brunel University London

2025

Acknowledgements

I want to express my deepest gratitude to all those who supported me throughout the journey. First and foremost, I would like to thank my supervisor, Dr Raha Pazoki, for the invaluable guidance, encouragement, and constructive feedback at every stage of my research. I am so grateful for your support, advice, and honest discussions throughout this experience. I would also like to thank Dr Cristina Sisu and Dr Ashley Houlden for their support and guidance during this journey.

I am also grateful to the College of Health, Medicine and Life Sciences, Brunel University of London, and all the staff for providing me with the resources and a supportive environment to carry out my research.

A special thanks goes to my colleagues Chukwueloka Hezekiah, Felix O'Farrell, Xijun (Hannah) Jiang and my friends who encouraged me, shared ideas, and stood by me during difficult moments. Finally, I am deeply thankful to my family, whose steady love, patience, and constant support gave me the strength and determination to complete this journey.

Abstract

Stroke is one of the leading causes of death and disability worldwide, with hypertension being a major risk factor for stroke. According to the World Stroke Organisation Global Stroke Fact Sheet, hypertension alone is responsible for over half of all stroke-related deaths and disability-adjusted life years (DALYs). While conventional risk factors for both diseases are well established, the added prediction value of genetic liability remains less clear.

The traditional risk prediction models for hypertension and stroke, such as the Framingham Hypertension Risk Score (FHRS) and the Framingham Stroke Risk Score (FSRS), typically rely on clinical and demographic factors, often assume linear effects, and typically overlook genetic liability and complex interactions between predictors.

In this thesis, three complementary studies were conducted to examine whether genetic liability could enhance the classification of hypertension and the prediction of strokes via both machine learning (ML) and traditional modelling techniques using data from more than 116,000 participants with European ancestry in the UK Biobank. Genetic variants and their effects obtained from genome-wide association studies were used to construct genetic liabilities for selected cardiovascular disease (CVD) risk factors and stroke, respectively. Multiple predictive models, such as Cox proportional hazards, penalized regression models (both logistic and Cox), tree-based algorithms (random forest, gradient boosting, and decision trees), and neural networks, were assessed after participants were randomly divided into training and testing sets.

Discrimination (AUC), calibration, and reclassification indices (NRI, IDI, and Brier score) were used to evaluate the models. Incorporating the genetic liabilities resulted in modest but steady improvements across the studies. The genetic liabilities associated with lipids improved the classification of

hypertension best with AUC using random forest. Additionally, stroke genetic liability enhanced stroke prediction with the Cox models, outperforming machine learning models. Among hypertensive individuals, the model's predictive performance (AUC) was higher in men and older adults than in women or younger adults. The Cox models outperformed all the machine learning models. Though ML methods allow for the investigation of non-linearities and interactions. Overall, genetic liability slightly enhances classification and risk prediction.

List of Figures

List of Abbreviations

AUC	Area Under Curve
ADS	Antioxidant Defence System
ANS	Autonomic Nervous System
BMI	Body mass index
CI	Confidence interval
CNV	Copy Number Variations
CVD	cardiovascular diseases
DF	Degrees of freedom
DNA	Deoxyribonucleic Acid
DZ	Dizygotic
FHRS	Framingham Hypertension Risk Score
FSRS	Framingham Stroke Risk Score
GWAS	Genome-wide association study
HRC	Haplotype Reference Consortium
HDL	High-density lipoprotein
IDI	Integrated Discrimination Improvement
INDEL	Insertion or Deletion
LD	Linkage disequilibrium
LDL	Low-density lipoprotein
ML	Machine learning
MZ	Monozygotic

NRI	Net Reclassification Index
RAS	Renin-Angiotensin System
ROS	Reactive Oxygen Species
SNPs	Single-nucleotide polymorphisms
SNS	Sympathetic Nervous System
T2DM	Type 2 diabetes mellitus
UK	United Kingdom

Table of Contents

Chapter One. General Introduction

1.1 Background

According to the World Health Organization, stroke and hypertension affect more than one billion individuals worldwide. Both conditions contribute significantly to the global economic and health burdens [1]. Non-modifiable risk factors such as sex, age, genetics and modifiable risk factors such as poor diet, sedentary lifestyle, diabetes, high cholesterol, and obesity contribute to the risk of both hypertension and stroke. Hypertension, which is a common condition in which the arterial blood pressure is consistently equal to or higher than 140/90 mmHg, is considered a major modifiable risk factor for stroke. Several epidemiological studies have demonstrated that untreated hypertension increases the risk of stroke by 2 to 4-fold [2]. The risk of stroke in hypertensive patients increases dramatically with additional risk factors such as obesity, diabetes or high cholesterol [3].

Developing models to accurately predict the risk of diseases such as stroke is crucial, particularly for high-risk populations like hypertensive individuals. Timely interventions like lifestyle adjustments and medical treatment will be possible if high-risk individuals are identified early. The incidence of stroke may significantly decline as a result.

The two well-known risk prediction models for stroke and hypertension are the Framingham Stroke Risk Score (FSRS) and the Framingham Hypertension Risk Score (FHRS). They both rely on simple statistical techniques.

Due to the availability of a large and complex health dataset, researchers are increasingly using artificial intelligence (AI) techniques, such as machine learning (ML) algorithms to develop risk prediction models to evaluate individuals' risk of a disease.

1.1.1 Pathophysiology of Hypertension

Hypertension is one of the established chronic health conditions globally and an important modifiable risk factor for cardiovascular diseases, including stroke, in the general population [4]. Many factors, including genetic predisposition,

lifestyle, obesity, insulin resistance, endothelial dysfunction, the renin-angiotensin system (RAS), and the sympathetic nervous system (SNS), contribute to the development of hypertension [5, 6].

A detailed description of the mechanisms involved in the development of hypertension has been thoroughly described previously in [5, 6]. Here, we describe the major processes that lead to hypertension. The process of developing hypertension is complex. Multiple physiological systems are involved in the process of developing hypertension. The renin-angiotensin system (RAS), the autonomic nervous system (ANS), and endothelin are all parts of these systems [5-9]. Two ways in which the ANS, especially the Sympathetic Nervous System (SNS), and the RAS interact in the circulatory system. First, angiotensin II (Ang II) raises SNS activity and oxidative stress, which causes blood vessels to narrow (vasoconstriction), the heart rate to rise, and retention of sodium. Second, the activation of the SNS stimulates the kidney to produce renin, increasing the RAS mechanism and increasing the production of Ang II in the system. This reciprocal relationship creates a feedback loop that raises blood pressure. Hypertension develops as a result of the aggravation of these two mechanisms [5-9].

Hypertension can lead to stroke through multiple processes. Chronic and uncontrolled hypertension can cause a variety of changes within the cardiovascular system, including cerebral circulation. These changes, including vascular remodeling, inflammation, oxidative stress, and baroreflex dysfunction, can lead to the development of stroke in hypertensive patients [4]. The key process is hypertension-induced oxidative stress, leading to an increase in reactive oxygen species (ROS) levels compared to the antioxidant defence system (ADS) in the brain and blood vessels. This excess ROS causes damage to the blood vessels and the brain, contributing to cerebrovascular dysfunction, characterised by endothelial damage, vascular stiffening, and

decreased blood flow. Any change in these systems can lead to atherosclerosis and increase the risk of stroke [4].

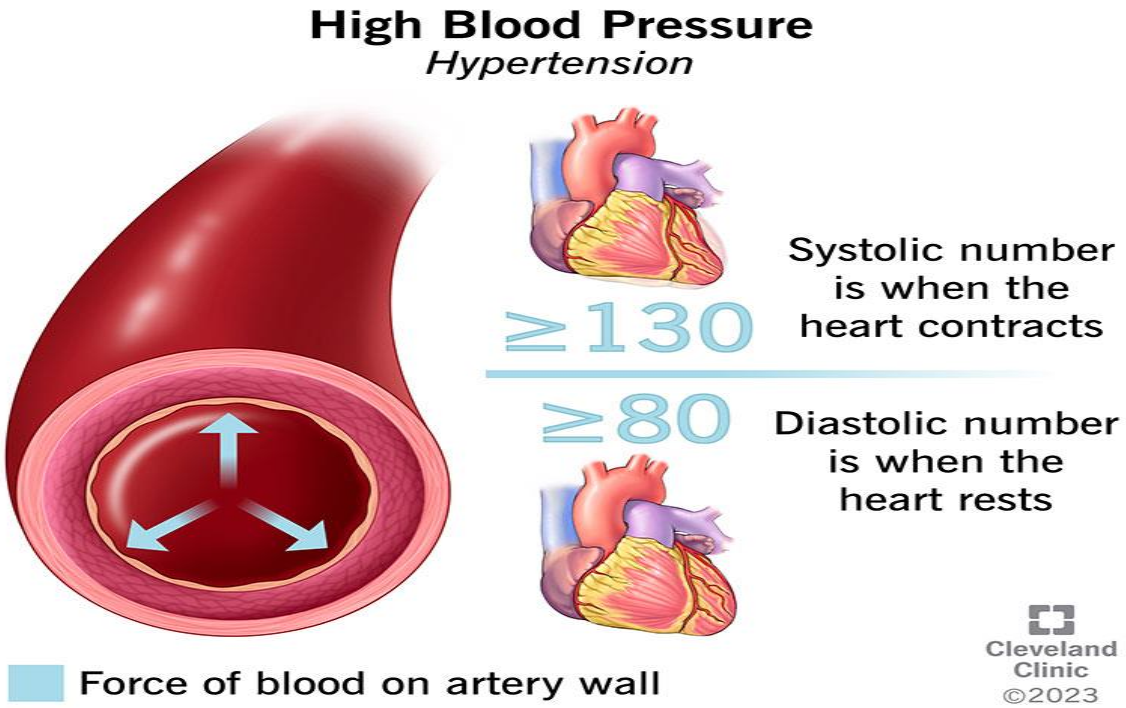


Figure 1.1. Illustration of the definition and understanding of hypertension.

1.1.2 Pathophysiology of Atherosclerosis

Atherosclerosis is a chronic inflammatory disease of the arteries [10], and it is one of the major causes of stroke. It is characterised by the formation of fatty plaques, including lipids, inflammatory cells, and fibrous elements, within the arterial walls [11]. The initial step of the pathophysiology of atherosclerosis development is endothelial injury caused by high blood pressure or high cholesterol levels, smoking, and diabetes. The injured endothelium becomes more permeable, allowing LDL cholesterol and other lipids to infiltrate the arterial wall. The LDL and other lipids are oxidised within the arterial wall, prompting inflammatory responses [11, 12]. The oxidised LDL and circulating monocytes move to the damaged endothelium site, forming cells, where the

monocytes attach to the endothelium and migrate into the arterial wall, differentiating into macrophages [12, 13].

The accumulation of foam cells, lipids, extracellular matrix, and other inflammatory cells leads to the growth of atherosclerotic plaques in the arterial walls. These plaques cause the arteries to narrow and harden, leading to disruption or restriction of blood flow and reducing oxygen supply to vital organs, including the brain [11, 12, 14]. Atherosclerosis can lead to stroke in numerous ways. As these plaques grow, they can become unstable and rupture, leading to the formation of a blood clot in the arteries [11].

These blood clots can completely block the artery, cutting off the blood supply and causing thrombotic stroke, or a piece of blood clot breaks off and travels through the bloodstream and lodges in a small artery in the brain, causing embolic stroke [12]. Essentially, both thrombotic and embolic strokes are ischemic strokes caused by clots blocking oxygenated blood from flowing to the brain.

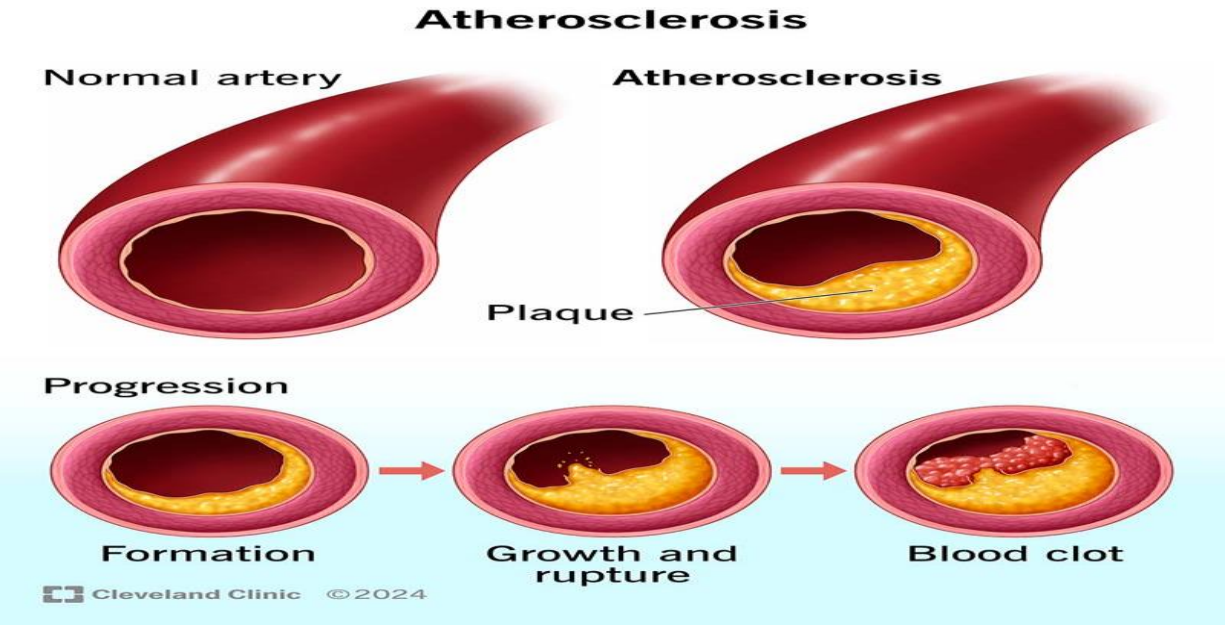


Figure 1.2. Illustration of the mechanism of plague formation and atherosclerosis progression: From a normal blood vessel to a partly blocked blood vessel due to plaque buildup [[15].

1.1.3 The Interplay of Hypertension and Atherosclerosis

Studies have suggested that atherosclerosis and hypertension each may enhance the oxidative stress of the arterial wall, and that the bridge connecting these conditions may be vascular inflammation [16-18]. Atherosclerosis and hypertension are not independent pathological conditions. They work together to promote the development of stroke. Hypertension accelerates atherosclerosis by causing damage to the arterial wall, promoting endothelial dysfunction, vascular inflammation, and lipid infiltration.

These processes can lead to plaque formation, causing the small cerebral arteries to become thick and narrow [18]. On the other hand, atherosclerosis promotes hypertension by causing the arterial walls to become narrower, decreasing vascular elasticity, i.e., increasing stiffness, and making it difficult for the heart to pump blood into the circulation system, which can result in hypertension [19-21]. This reciprocal relationship between hypertension and atherosclerosis can make the cerebrovascular system weak and increase the risk of stroke.

1.1.4 Stroke

Stroke is the second leading cause of death and a major cause of adult disability worldwide [22, 23]. It is a medical emergency caused by the blockage or rupture of blood vessels in the brain. It occurs when the blood supply to the brain is interrupted by a rupture of the blood vessel in the brain or reduced by a clot formed in an artery, causing a blockage of a blood vessel and cutting off oxygen to the brain, which leads to the death of brain cells.

According to the UK Stroke Association, cerebrovascular disease is the fourth most common cause of death in the United Kingdom, and stroke, one of the cerebrovascular diseases, causes around 38,000 deaths each year in the UK. Stroke accounts for about 75% of deaths from cerebrovascular diseases. There are around 100,000 strokes every year, and approximately 1.4 million people are living with stroke in the UK. Two major types of strokes, ischemic and

hemorrhagic, result from distinct mechanisms, and their causes are diverse and interconnected. The ischemic stroke, which accounts for more than 85% of all strokes, is caused by a blockage or blood clot in a brain artery, which can occur as a result of a thrombotic condition in which blood flow to the brain is affected by vessel hardening or narrowing due to atherosclerosis, or an embolic condition in which a blood clot formed in another part of the body, travels to the brain and blocks a smaller artery [23].

Hemorrhagic stroke is caused by rupture of the blood vessels in the brain, is responsible for about 15% of all strokes, and has a high mortality rate. It is classified into intracerebral haemorrhage, where the rupture occurs within the brain tissue due to stress in the brain tissue, and internal injury caused by chronic hypertension and subarachnoid haemorrhage, where blood accumulates in the subarachnoid space of the brain due to a head injury or ruptured cerebral aneurysm [23]

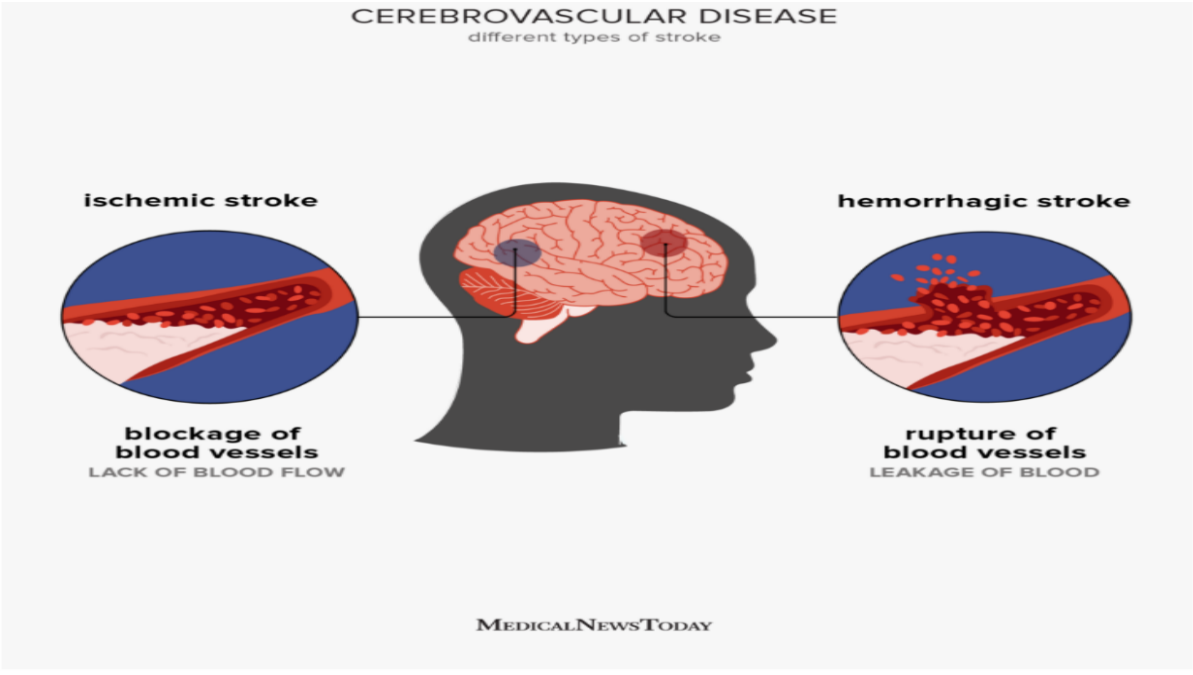


Figure 1.3 Illustration of two main types of strokes and their causes. Recreated by Yaha Mulcare.

1.1.5 Genetic Factors for Complex Diseases

The causes of complex diseases such as stroke, hypertension, and many more involve multiple factors, including genetic predisposition, environmental exposures, lifestyle, and the social and economic context [24].

Family and twin studies have provided evidence on the role of genetic factors in complex diseases. The results from family studies confirm a significant association between family history and stroke risk, indicating heritability. Parental stroke was observed to be associated with a 3-fold increase in risk of stroke in their children [25]. For example, men whose mothers died due to stroke have a 3-fold increased incidence of stroke in comparison with men without a maternal history of stroke. Also, stroke in a first-degree relative increases an individual's odds by 2- to 6-fold [26].

Studies have shown that a family history of ischemic stroke increases the risk of ischemic stroke, while a family history of hemorrhagic stroke increases the risk of hemorrhagic stroke [[27]. For example, if one sibling experienced a haemorrhagic stroke, the risk of haemorrhagic stroke for the other sibling was reported to be 2.14. The corresponding risk for ischemic stroke was 1.82. The overall risk for all types was reported to be 1.67 [27]. This suggests that ischemic and haemorrhagic stroke do not share the same genetic influence.

Twin studies are considered more informative than traditional family studies because they can separate the effects of genetic factors and the common environment. These studies have been instrumental in determining and understanding the genetic basis for stroke and other complex diseases [28, 29]. Further evidence of the role of genetic factors in stroke risk has come from twin studies showing that monozygotic (MZ) twins have a 2- to 4-fold increase in stroke risk compared to dizygotic (DZ) twins. Twin studies use the ratio of disease concordance between MZ and DZ twins to estimate the most likely effects of genetics [28]. These findings demonstrate the significant influence of genetic predisposition on stroke susceptibility and underline the importance of

incorporating genetic models into the study of cerebrovascular disease. In addition to improving our knowledge of stroke pathophysiology, these models open the door for precision medicine strategies like targeted therapies, early intervention, and risk stratification.

1.1.6 Overview of Genomic Variations

The human genome is a complete set of genetic instructions encoded in the Deoxyribonucleic acid (DNA) of humans. It is made up of more than three billion base pairs of DNA and is arranged into 23 pairs of chromosomes [30, 31]. DNA instructions are present in every cell of the human body, made up of four different chemicals. They are called nucleotides or bases, and each one has its own letter. Adenine (A), thymine (T), cytosine (C), and guanine (G). There are minor differences in DNA sequences among people within a population. This is referred to as genome variation. These variations, also known as variants, contribute to most of the genetic diversity within a population and greatly contribute to disease susceptibility among individuals in the population [32, 33].

The three main types of genome variation are single-nucleotide polymorphisms (SNPs), insertions or deletions (INDELs), and structural variation (SV). The SV includes both copy number variation (CNV) and chromosomal rearrangement events, such as inversions and translocations [[34]. SNPs and INDELs account for most inherited phenotypes, including disease susceptibility.

1.1.7 Single Nucleotide Polymorphism (SNPs)

A SNP is a change in a single base (nucleotide position, such as A to G) in the genome that occurs at a frequency greater than 1% in a population.

SNPs are the most popular type of variation in the human genome, and they are useful tools for a wide range of genetic studies. SNPs are regarded as the most useful biomarkers in genome-wide association studies (GWAS) that connect loci to complex diseases [35, 36].

1.1.8 Genome-Wide Association Studies (GWAS)

GWAS is a method in which hundreds of thousands to millions of genetic variants (SNPs) across the genomes are evaluated to discover those that are statistically associated with a specific disease or trait [[37]. The most common GWAS is the case-control format, in which two large groups of individuals, one healthy group(controls) and one diseased group (cases), are compared. In GWAS, SNP allele frequencies, which indicate how common different variants occur in the population, are compared between cases and controls to discover SNPs that are statistically associated with specific diseases. The evidence derived from GWAS may be considered robust, as it does not depend on the accuracy of a prior functional hypothesis. These SNPs could explain gene causality in complex diseases [38].

Over 6000 GWAS studies and more than 500,000 SNP-trait associations have been published in the publicly available NHGRI-EBI GWAS Catalog. These studies have discovered hundreds of thousands of SNP-disease associations, transforming our understanding of the genetic basis of complex diseases [39]. However, most of the SNPs have small effect sizes and explain only a small proportion of heritability [40, 41]. For example, a multi-ancestry GWAS has identified 32 loci associated with stroke and stroke subtypes [42, 43]. Studies have used GWAS results and revealed a significant association between genetic variants associated with risk factors of stroke, such as diabetes, obesity, etc., and the risk of hypertension and increased risk of CVD in the general population.

1.1.9 Genetic liability: Polygenic Risk Score (PRS)

It is documented that the genetic liability calculated based on identified risk alleles is predictive of the incidence of disease and provides a continuous and quantitative measure of genetic susceptibility [44]. Both hypertension and stroke are polygenetic diseases, with many risk alleles of small effect. A genetic liability approach allows the cumulative effect of multiple genetic risk variants

on disease status to be analysed, clarifying whether a genetic predisposition is associated with the disease.

Several studies have reported that a genetic liability derived from a set of SNPs associated with stroke or its risk factors has a significant association with stroke risk [[45-51]. For instance, studies have demonstrated that genetic liability derived from stroke GWAS meta-analysis is independently associated with risk of stroke incidents, and among individuals with high genetic liability, lifestyle factors had the most influence on risk [46, 47].

Genome-wide genetic liabilities, derived from the combined effects of several genetic variants across the genome, regardless of the strength of their association, have been increasingly tested in recent years for their effects on health and disease, and previous studies have shown that higher scores of genome-wide genetic liabilities enhance stroke risk prediction [52, 53].

1.2 Research Problem and Gaps

The Framingham Stroke Risk Score (FSRS) and the Framingham Hypertension Risk Score (FHRS) are two well-known examples of traditional cardiovascular risk prediction methods designed to estimate the risks of stroke and hypertension, respectively. These models were created using standard statistical techniques, relying mainly on clinical and demographic factors, often assuming linear effects, and typically overlooking genetic liability and complex and non-linear interactions between the predictors. Here, three complementary studies were conducted utilizing the UK Biobank data to assess the incremental value of genetic liability in disease classification and prediction via both machine learning and traditional modelling techniques.

Although several studies have investigated the risk and predictors of stroke and hypertension and demonstrated that including genetic liability in epidemiological models improves precision for both stroke and hypertension risk when compared to models without genetic liability [47, 54-56]. There is

little or insufficient evidence about the added prediction value of integration of genetic liability or multiple genetic liabilities into risk prediction models, especially for subgroups. These limitations are more noticeable when considering high-risk groups, particularly hypertensive patients, hypertensive men versus hypertensive women, and older versus younger hypertensive patients in the context of large-scale cohort research such as the UK Biobank. Addressing these limitations is important for developing population-specific risk prediction models that can improve stroke risk assessment, preventative measures, and policies for hypertensive patients.

Several studies have applied both machine learning and traditional statistical techniques and compared their prediction ability in the prediction of stroke, but the findings have been inconsistent. Some studies have reported superior performance of machine learning models over traditional statistical models; others have found that traditional statistical methods performed similarly well or better. This inconsistency highlights the need for more investigation into the relative advantages of each approach in different populations and clinical settings, as well as the need to develop new machine learning approaches with enhanced prediction value.

This thesis aims to address these limitations by incorporating stroke genetic liability into both machine learning and traditional statistical prediction frameworks, allowing for the inclusion of stroke genome-wide genetic predisposition in individual stroke risk evaluations.

Therefore, the objective of this thesis is to evaluate and compare the predictive value of genetic liability-integrated machine learning and traditional statistical models for hypertension and stroke. Additionally, assessing the predictive value of genetic liability-integrated machine learning and traditional statistical models for stroke risk in hypertensive individuals within the UK Biobank, with subgroup analyses by sex and age group to inform targeted prevention policies.

1.2.1 Aims and Objectives

The central goal of this thesis is to evaluate the added predictive value of genetic liability in the risk prediction models for risk of disease, with a focus on three main objectives:

1. Assessing the predictive value of multiple genetic liabilities to CVD risk factors in the classification of hypertension (Chapter three)
2. Assessing the predictive value of stroke genetic liability in the prediction of stroke (Chapter four)
3. Assessing the predictive value of stroke genetic liability in the prediction of stroke in hypertensive individuals (Chapter five)

We addressed these objectives through three research studies, each presented as a separate manuscript in chapters three, four and five.

1.2.2 Structure of the Thesis

This thesis has been organised in six chapters and composed of three published or submitted core manuscripts, which are grouped into Introduction, Methods, Results, and Discussion sections to reflect the main aim and the research objectives of this study. Each results section in chapters three, four, and five addresses a specific research objective.

Chapter One introduces and outlines the research by providing the background and context, including the purpose and objectives of the research. It contextualizes the research within the broader context of predicting cardiovascular disease, as well as highlighting the significance and rationale of the research.

Chapter two presents the overview of the research methodology, which includes the conceptual framework, data sources, and data collection strategies. The chapter also provides descriptions of the analysis and the modeling techniques in the analysis, including machine learning and statistical methods.

Chapter three addresses the first objective of the research, which is to explore whether the genetic liability to multiple cardiovascular risk factors can improve hypertension classification through the application of machine learning methods.

Chapter Four addresses the second research objective, which assessed whether or not the stroke genetic liability improved the prediction of stroke risk in the cohort of interest. This chapter evaluates and directly compares the relative performance of both machine learning models and traditional statistical methods.

Chapter 5 modifies the analysis by focusing on the high-risk subgroup, the hypertensive patients. This chapter considers the role of the stroke genetic liability in the predictive performance of strokes in this group. Moreover, the chapter examines whether the predictive performance of strokes in the group varies by age and sex.

Chapter Six presents a general discussion and summary of the main findings of the research. It outlines the implications for disease risk prediction and early intervention, while also mentioning several areas of research that could be explored to further improve prediction and classification accuracy.

1.2.3 Summary of Chapter One

The introductory chapter describes the layout of the thesis. It outlines the overall aim and research objectives, provides context and background of the research, and highlights the originality and contributions of the research. As it concludes, the chapter provides a summary of the structure of the thesis and presents a clear road map for the content and progression of the remaining chapters.

2 Chapter Two. Research Design and Methodology

This chapter outlines the general scope of the research. It describes the study population, data source, data collection methods, and various forms of analysis used, ranging from classical statistics to machine learning.

2.1 Study Population and Data Source

The thesis is based on a subset of unrelated individuals of European ancestry in the UK Biobank (UKB). The UKB project is a large prospective cohort study of approximately 500,000 participants from across the United Kingdom, aged between 40 and 69 during recruitment. The detailed description of the UKB project and SNP genotyping and imputation is provided in [57-61]. The UKB genotyped around 500,000 persons using two custom SNP arrays created with Affymetrix: the UK BiLEVE Axiom Array (about 50,000 samples) and the UK Biobank Axiom Array (about 450,000 samples). Both arrays overlap significantly and share more than 95% of SNP content, allowing for integrated analyses [58, 59]. Approximately 820,000 SNPs per sample were genotyped using the Affymetrix Axiom, followed by large-scale imputation with the Haplotype Reference Consortium (HRC) and combined UK10K and 1000 Genomes Phase 3 reference panels. The imputation involves using powerful software such as IMPUTES to infer missing genotypes in the microarray data. Following the imputation, more than 90 million variations (SNPs and INDELs) were catalogued in the latest dbSNP database [62].

To identify individuals of European descent, we performed k-means clustering analysis on the genetic principal component data created centrally by the UKB. We then obtained genetic data from the individuals who had passed the UKB internal quality control (QC) and had genotype data. UKB internal QC is described in detail by others and in [59]. Briefly, UKB performed two main quality controls (QC), the sample-based and marker-based. The sample-based QC, designed to identify samples with poor quality genotype calls, find related individuals and describes the ancestral diversity of the cohort based on genetic data, while the marker-based QC procedures are used to account for effects such as the large cohort size, population structure, and batch-based genotype calling. We excluded individuals with mismatched genetics and self-reported data to avoid potential inconsistencies in data reporting. Using the kinship cutoff of

0.0884 for third-degree relatives, we further excluded participants who were up to second-degree related.

Additionally, we excluded individuals who had been diagnosed with a stroke, heart attack, or angina before or at baseline. This strategy helps to adjust for pre-existing CVDs and minimizes the possibility of confounding. We excluded participants who had withdrawn their consent, pregnant individuals, or those who were uncertain about their pregnancy status.

Furthermore, individuals who were on cholesterol-lowering medication, stopped smoking or drinking due to health reasons or doctor's advice, and participants with missing data on the potential confounders were excluded from the dataset.

2.1.1 Definition of the Outcomes

When in taking blood pressure measure inaccurate readings can occur, often due to measurement methods or improper position of the patients. Therefore, using single reading to determine patient's hypertension status can lead to misclassification of hypertension category. To minimise misclassification, the American Heart Association (AHA) and American College of Cardiology (ACC) have recommended blood pressure (BP) averaging [63]. The process involves taking multiple BP readings and then take the average of the values. In chapter three, our main outcome is hypertension, which was defined as (1) the presence of a recorded average of SBP $\geq$ 140 mmHg or a DBP $\geq$ 90, or (2) hypertension diagnosed by a doctor, or (3) a record of using blood pressure-lowering medication at baseline. In chapters four and five, our primary outcome, incident stroke, was characterised using the cerebrovascular disorders International Classification of Diseases 10th revision (ICD-10, I60–I67), and the follow-up period is computed from the date of health assessment upon enrolment to the end of March 2017. The chapter further restricted the analysis to the prediction of stroke risk in hypertensive participants.

2.1.2 SNP Selection and Computation of Genetic Liabilities

The main exposures considered in this thesis are the genetic liabilities for CVD risk factors and stroke. From genetic liability, researchers can estimate an individual's risk of developing a particular disease based on their disease-susceptible genetic variants, which were discovered to be linked to the risk of the disease [64]. There are two main approaches to estimating genetic liability: unweighted and weighted methods. Based on the number of risk alleles carried by an individual, genetic liability is generated for each individual, and a score is assigned to everyone according to the number of risk alleles they carry. In the unweighted approach, the genetic liability of each participant is calculated by summing up the number of risk alleles (i.e., 0, 1, or 2) at each selected SNP. Thus, the genotype code "0" suggests no risk allele; the risk allele heterozygotes are coded as genotype "1," and risk allele homozygotes are coded as genotype "2".

In the weighted approach, instead of assigning equal weight to each SNP, the genetic liability is generated by multiplying the reported effect size (β-coefficients of the SNPs) and the number of risk alleles at the selected SNPs. and the product is summed or averaged to create the genetic liability for each participant in the study.

In the current thesis, the weighted GRS approach was selected over the unweighted version because it incorporates the effect size of each genetic variant, allowing variants with stronger associations with the trait or disease to contribute more to the score. This provides a more accurate reflection of genetic risk compared to treating all variants equally. The genetic liabilities were calculated using GWAS summary statistics that excluded UK Biobank variants. For the classification of the hypertension, multiple genetic liabilities were generated using a list of genetic variants in the form of SNPs (Supplementary Data S1–S10) that were previously identified to be associated with ten CVD

risk factors, including type 2 diabetes, two adiposity traits, three smoking traits, and four lipid traits at a GWAS significant threshold (p-value < 5.0 × 10−8) in the European population. The "- -score" function of PLINK with the sum option was employed for the generation of genetic liabilities for type 2 diabetes, adiposity traits, smoking traits, and lipid traits.

For stroke prediction, a single genetic liability for stroke was generated using variants associated with stroke from MEGASTROKE GWAS (**Supplementary Data S1**) using the "- - score" function with average and non-mean-imputation options in PLINK. The MEGASTROKE consortium is a large-scale international collaboration established by the International Stroke Genetics Consortium. They provide researchers access to the summary statistics of the 2018 meta-analysis GWAS data on stroke and stroke subtypes, enabling other researchers to explore these data for scientific purposes. The publicly available GWAS data on stroke are accessible via the MEGASTROKE consortium publication [42]. The effect sizes for all were obtained from GWAS summary statistics data that were published and made publicly available on the GWAS catalog website (**https://www.ebi.ac.uk/gwas/**, accessed on 12 July 2021). PLINK 1.9 was used to generate all the genetic liabilities after the necessary quality control processes were completed.

2.2 Statistical methods used in this thesis

We assessed the association between the genetic liabilities and outcomes of interest using both univariable and multivariable statistical methods. The logistic regression technique was used to evaluate the association of the genetic liabilities for ten CVD risk factors and risk of hypertension (prevalent hypertension), whereas the Cox regression approach was used to test the association of stroke genetic liability and risk of future stroke. In all models, we adjusted for age, sex, BMI, drinking, smoking, diabetes, and cholesterol level. All analysis was performed using the R program 4.2.2.

2.2.1 Statistical models (logistic, survival regression)

Traditional statistical models have long been the foundation of data analysis. They provide researchers with organised ways to discover relationships in data, make predictions, and draw meaningful conclusions about populations based on smaller samples. These models, however, rely on set assumptions about the data, and the strength of their results often depends on how well those assumptions are met [65, 66].

A major advantage of traditional statistical models is their clarity. With relatively small datasets, they are efficient, and the results are often straightforward to interpret, allowing researchers to understand precisely how variables relate to one another. However, they are inefficient with high-dimensional and complex data and may not be able to capture complex, non-linear relationship in the data [65, 66].

One of the most widely used traditional statistical models is regression analysis, which examines the relationship between a dependent variable (the outcome of interest) and one or more independent variables (the predictors of the outcome). In simple terms, regression helps us understand how changes in one factor are associated with changes in the outcome while keeping other factors constant. In the world of medical research, two regression techniques are prominent: The logistic regression and survival analysis (like the Cox regression or Cox proportional hazards model).

Logistic regression analysis is a statistical method used to determine the relationship between a binary or multinomial outcome and various explanatory variables, which can be either categories or continuous measurements. Instead of predicting an exact value, logistic regression estimates the probability of an event happening or its odds ratio [67].

The Cox Proportional Hazards Model (Cox Model) is a widely used regression method in survival analysis. It is primarily employed for examining risk factors associated with diseases, mortality, or other survival outcomes among different

patient groups [68]. A key assumption behind the Cox model is the 'proportional hazards assumption,' which states that the risk ratio between any two individuals remains constant over time [69]. This implies that the relative risk between two people does not change throughout the follow-up period. What makes the Cox model so flexible and useful for epidemiological studies is that, unlike many other methods, it does not require assumptions about the shape of the underlying baseline hazard function [68].

2.2.2 Discrete and continuous time survival models

Traditional methods for survival analysis typically treat time to event as a continuous outcome. The survival prediction problem is formulated as a censored regression problem, relying on a set of parametric or semi-parametric assumptions on the survival times [70, 71]. The continuous time survival models are designed for situations where the event of interest can occur at any point in time, which is useful when precise timing of events is available and relevant. The Cox Proportional Hazards (CoxPH) model is the most commonly used semi-parametric model, which relies on the proportional hazard assumption about survival times [70]. These models can be more efficient and accurate when the survival times follow the assumed parametric distribution, but when they are violated, they can result in inaccurate predictions [71].

An alternative approach to continuous-time survival analysis is a discrete-time survival model. In a discrete-time survival framework, the continuous survival time data is transformed into a discrete-time format, a set of time bins. This approach reformulates the survival analysis problem as a series of binary classification problems, which is a type of multi-task learning problem [70, 71]. This approach is useful where data is collected at regular intervals or time is naturally measured in discrete units or when the exact timing of an event within an interval is unknown, but it is known to have occurred within that interval. The discrete-time survival models often leverage standard regression

techniques, particularly logistic regression, and do not assume proportional hazards across intervals in the same way continuous models do, making them more flexible when the proportional hazards assumption is violated.

2.2.3 Machine learning models

Machine Learning (ML) is a branch of Artificial Intelligence (AI) where systems utilise algorithms to learn from data, identify patterns, and make decisions with minimal human intervention. ML algorithms are trained on datasets, allowing them to discover underlying relationships and improve their performance over time [66]. One main difference between ML and traditional statistical methods lies in their ability to make predictions as accurately as possible, while the latter are aimed at inferring relationships between variables [66, 71]. Several ML models can be trained on a large dataset and make accurate predictions. In this thesis, the ML models applied include random Forest (RF), Gradient Boosting Machine (GBM), Neural Network (Nnet), Decision Tree (DT), and penalized models (Glmnet). Detailed description and implementation of these models have been explained in the subsequent chapters. The Decision Tree is one of the common and simple methods used for classification and regression applications. It works by dividing a dataset into smaller subgroups depending on feature values and then generating a decision tree. Random Forest is a popular machine learning model for classification and regression. It creates ensembles from decision trees and combines their results to make a final decision. The Gradient Boosting Machine models integrate predictions from many weak learners to increase total prediction accuracy. The neural network consists of interconnected processing nodes organized in three layers: the input, hidden, and output layers. To estimate generalisation performance and to select the best performing parameters as well as to evaluate the reliability of ML model performances, we applied the k-fold cross validation (CV) technique to evaluate the best performing values for hyperparameters. In this thesis, we implemented 10-fold CV to tune the hyperparameters. The process involves partitioning the training set into 10 subsets, each subset containing an approximately equal size of the data. The model is trained and tested 10 rounds, using a different subset as testing set and the rest as training set in each round. The process ends up with 10 testing scores, one per subset. The model performance is assessed across different hyperparameter values and the value that generates best model performance metric is selected. In this

thesis, several performance metrics was used. The hyperparameter with the smallest CV root mean square error (cv-rmse) was used to develop the Glmnet models and GBM whereas CV error (x-error), the receiver operation characteristic(ROC) and Out-of-Bag (OOB) error were utilized to develop DT, Nnet and RF prediction models, respectively.

The performance criteria used in the papers (chapters three, four, and five) are summarised in the table below

Table 1. Hyperparameter tuning performance criteria used in the published and submitted papers

performance criteria	Used in Paper	Models	Description
OOB error	1, 2, 3	RF	The OOB error serves as an important metric for assessing the generalisation performance of a random forest model on unseen data. Its estimation is helpful in the selection of optimal hyperparameters values during the model tuning process. It is calculated using out-of-bag samples, which are defined as the data points from the original training set that are not used in the development of the random forest ensemble.
ROC	1, 3	Nnet	ROC(AUC) assesses the binary model's ability to distinguish between cases and non-cases across all possible classification thresholds, serving as a key performance metric in hyperparameter tuning process for binary neural network classification. Higher values indicate better performance.
x-error	2	DT	The x-error measures the model's ability for generalisation to unseen data, that was not used during the model training stage. It is commonly estimated through k-fold cross-validation. The x-error is instrumental in the selection of the best hyperparameters and model complexity parameter (cp) during the pruning process.
cv-rmse	2, 3	GMB, Glmnet	The cv-rmse is a critical metric for assessing a model’s predictive performance and its ability to generalise to unseen data. It plays significant role in the selection the best hyperparameters for GBM and Glment models. The hyperparameters that yield lowest cv-rmse are used to develop the final predictive models.

The model performance and added prediction value of genetic liability were assessed based on discrimination ability (area under the curve), net reclassification index (NRI), and integrated discrimination improvement (IDI). The detailed description of the implementation of these models has been presented in the three published or submitted manuscripts (chapters three, four, and five).

2.2.4 Summary of Chapter Two

In this chapter, the research methodology and procedure followed in the study were described and explained. The use of UKB data, genetic variants from GWAS to generate genetic liability, statistical and machine learning are used for analysing the data. Finally, describe how the model performance and added prediction value of genetic liability were assessed. The following chapter will present the descriptive findings of the study.

References

1. Cheng Y, Lin Y, Shi H, Cheng M, Zhang B, Liu X, Shi C, Wang Y, Xia C, Xie W: **Projections of the Stroke Burden at the Global, Regional, and National Levels up to 2050 Based on the Global Burden of Disease Study 2021.** Journal of the American Heart Association 2024, **13**(23):e036142.

2. Magid-Bernstein J, Girard R, Polster S, Srinath A, Romanos S, Awad IA, Sansing LH: **Cerebral Hemorrhage: Pathophysiology, Treatment, and Future Directions.** Circulation research 2022, **130**(8):1204–1229.

3. Du X, McNamee R, Cruickshank K: **Stroke Risk from Multiple Risk Factors Combined with Hypertension: A Primary Care Based Case-control Study in a Defined Population of Northwest England.** Annals of epidemiology 2000, **10**(6):380–388.

4. Yu J, Zhou R, Cai G: **From Hypertension to Stroke: Mechanisms and Potential Prevention Strategies.** CNS neuroscience & therapeutics 2011, **17**(5):577–584.

5. Beevers G: **ABC of hypertension: The pathophysiology of hypertension.** BMJ 2001, **322**(7291):912–916.

6. Harrison DG, Coffman TM, Wilcox CS: **Pathophysiology of Hypertension: The Mosaic Theory and Beyond.** Circulation research 2021, **128**(7):847–863.

7. Miller AJ, Arnold AC: **The renin–angiotensin system in cardiovascular autonomic control: recent developments and clinical implications.** Clin Auton Res 2019, **29**(2):231–243.

8. de Champlain J, D'Orléans-Juste P: **Role of the renin-angiotensin system on the central and peripheral autonomic nervous system.** In *ACE Inhibitors.* Edited by Anonymous Switzerland: Springer Basel AG; 2001:145–153.

9. Dong T, Chen J, Tian L, Wang L, Jiang R, Zhang Z, Xu J, Zhao X, Zhu W, Wang G, Sun W, Zhang G: **Role of the Renin-Angiotensin System, Renal Sympathetic Nerve System, and Oxidative Stress in Chronic Foot Shock-Induced Hypertension in Rats.** International journal of biological sciences 2015, **11**(6):652–663.

10. Galkina E, Ley K: **Immune and Inflammatory Mechanisms of Atherosclerosis.** Annual Review of Immunology 2009, **27**(1):165–197.

11. Jebari-Benslaiman S, Galicia-García U, Larrea-Sebal A, Olaetxea JR, Alloza I, Vandenbroeck K, Benito-Vicente A, Martín C: **Pathophysiology of Atherosclerosis.** International Journal of Molecular Sciences 2022, **23**(6):3346.

12. Bergheanu SC, Bodde MC, Jukema JW: **Pathophysiology and treatment of atherosclerosis.** Neth Heart J 2017, **25**(4):231–242.

13. Tousoulis D, Kampoli A, Papageorgiou N, Androulakis E, Antoniades C, Toutouzas K, Stefanadis C: **Pathophysiology of Atherosclerosis: The Role of Inflammation.** Current pharmaceutical design 2011, **17**(37):4089–4110.

14. Tsivgoulis G, Safouris A, Kim D, Alexandrov AV: **Recent Advances in Primary and Secondary Prevention of Atherosclerotic Stroke.** Journal of stroke 2018, **20**(3):417.

15. Bentzon JF, Otsuka F, Virmani R, Falk E: **Mechanisms of Plaque Formation and Rupture.** Circulation Research 2014, **114**(12):1852–1866.

16. Li J, Chen J: **Inflammation may be a bridge connecting hypertension and atherosclerosis.** Medical hypotheses 2005, **64**(5):925–929.

17. Siti HN, Kamisah Y, Kamsiah J: **The role of oxidative stress, antioxidants and vascular inflammation in cardiovascular disease (a review).** Vascular Pharmacology 2015, **71**:40–56.

18. Cachofeiro V, Miana M, Heras N, Martín-Fernandez B, Ballesteros S, Balfagon G, Lahera V: **Inflammation: A Link Between Hypertension and Atherosclerosis.** Current hypertension reviews 2009, **5**(1):40–48.

19. Kim H: **Arterial stiffness and hypertension.** Clin Hypertens 2023, **29**(1):31–9.

20. Payne RA, Wilkinson IB, Webb DJ: **Arterial Stiffness and Hypertension: Emerging Concepts.** Hypertension 2010, **55**(1):9–14.

21. Nakano H, Shiina K, Takahashi T, Fujii M, Iwasaki Y, Matsumoto C, Yamashina A, Chikamori T, Tomiyama H: **Bidirectional Longitudinal Relationships Between Arterial Stiffness and Hypertension Are Independent of Those Between Arterial Stiffness and Diabetes: A Large-Scale Prospective Observational Study in Employees of a Japanese Company.** Journal of the American Heart Association 2022, **11**(13):e025924.

22. Sarah Rizqiya Zahiya Muharrika Dias Syukriyah, Asra Al Fauzi: **Hypertension as a risk factor in stroke: An overview.** World Journal of Advanced Research and Reviews 2024, **21**(1):2370–2372.

23. Kuriakose D, Xiao Z: **Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives.** International Journal of Molecular Sciences 2020, **21**(20):7609.

24. Mitchell KJ: **What is complex about complex disorders?** Genome Biology (Online Edition) 2012, **13**(1):237–2717.

25. Seshadri S, Beiser A, Pikula A, Himali JJ, Kelly-Hayes M, Debette S, DeStefano AL, Romero JR, Kase CS, Wolf PA: **Parental Occurrence of Stroke and Risk of Stroke in Their Children.** Circulation (New York, N.Y.) 2010, **121**(11):1304–1312.

26. Huang H, Yang C, Shu H, Kuang Y, Yang T, He W, Zhao K, Xia X, Cheng J, Ma Y, Gu J: **Genetic predisposition of stroke: understanding the evolving landscape through meta-analysis.** International journal of clinical and experimental medicine 2015, **8**(1):1315–1323.

27. Sundquist K, Li X, Hemminki K: **Familial Risk of Ischemic and Hemorrhagic Stroke: A Large-Scale Study of the Swedish Population.** Stroke 2006, **37**(7):1668–1673.

28. Bak S, Gaist D, Sindrup SH, Skytthe A, Christensen K: **Genetic Liability in Stroke: A Long-Term Follow-Up Study of Danish Twins.** Stroke 2002, **33**(3):769–774.

29. BAIRD AE: **Genetics and Genomics of Stroke: Novel Approaches.** Journal of the American College of Cardiology 2010, **56**(4):245–253.

30. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** Nature 2010, **467**(7319):1061–1073.

31. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST,

Wang J, Wilson RK, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS: **A global reference for human genetic variation.** Nature 2015, **526**(7571):68–74.

32. Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits.** Nat Rev Genet 2005, **6**(2):95–108.

33. Rebbeck TR, Spitz M, Wu X: **Assessing the function of genetic variants in candidate gene association studies.** Nat Rev Genet 2004, **5**(8):589–597.

34. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS: **The discovery of human genetic variations and their use as disease markers: past, present and future.** J Hum Genet 2010, **55**(7):403–415.

35. Anonymous **10 Years of GWAS Discovery: Biology, Function, and Translation.**

36. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.** Nucleic Acids Research 2014, **42**(D1):1001.

37. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D: **Genome-wide association studies.** Nat Rev Methods Primers 2021, **1**(1):59.

38. Momozawa Y, Mizukami K: **Unique roles of rare variants in the genetics of complex diseases in humans.** Journal of Human Genetics 2021, **66**(1):11–23.

39. Tam CHT, Lim CKP, Luk AOY, Ng ACW, Lee H, Jiang G, Lau ESH, Fan B, Wan R, Kong APS, Tam W, Ozaki R, Chow EYK, Lee K, Siu S, Hui G, Tsang C, Lau K, Leung JYY, Tsang M, Kam G, Lau I, Li JKY, Yeung VTF, Lau E, Lo S, Fung S, Cheng Y, Chow C, Hu M, Yu W, Tsui SKW, Huang Y, Lan H, Szeto C, Tang NLS, Ng MCY, So W, Tomlinson B, Chan JCN, Ma RCW: **Development of genome-wide polygenic risk scores for lipid traits and clinical applications for dyslipidemia, subclinical atherosclerosis, and diabetes cardiovascular complications among East Asians.** Genome medicine 2021, **13**(1):29.

40. Choi S, Bae S, Park T: **Risk Prediction Using Genome-Wide Association Studies on Type 2 Diabetes.** Genomics & informatics 2016, **14**(4):138–148.

41. Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, Yengo L, Lloyd-Jones LR, Sidorenko J, Agbessi M, Ahsan H, Alves I, Andiappan AK, Awadalla P, Battle A, Beutner F, Bonder M, Boomsma D, Christiansen M, Claringbould A, Deelen P, Esko T, Favé M-, Franke L, Frayling T, Gharib S, Gibson G, Hemani G, Jansen R, Kähönen M, Kalnapenkis A, Kasela S, Kettunen J, Kim Y, Kirsten H, Kovacs P, Krohn K, Kronberg-Guzman J, Kukushkina V, Kutalik Z, Lee B, Lehtimäki T, Loeffler M, Marigorta U, Metspalu A, Milani L, Müller-Nurasyid M, Nauck M, Nivard M, Penninx B, Perola M, Pervjakova N, Pierce A, Powell J, Prokisch H, Psaty B, Raitakari O, Ring S, Ripatti S, Rotzschke O, Ruëger S,

Saha A, Scholz M, Schramm K, Seppälä I, Stumvoll M, Sullivan P, Teumer A, Thiery J, Tong L, Tönjes A, Dongen J, Meurs J, Verlouw J, Völker U, Võsa U, Yaghootkar H, Zeng B, McRae AF, Visscher P, Zeng J, Yang J: **Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes.** Nature communications 2018, **9**(1):2941–14.

42. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese AK, van Der Laan SW, Gretarsdottir S, Anderson CD, Chong M, Adams HHH, Ago T, Almgren P, Amouyel P, Ay H, Bartz TM, Benavente OR, Bevan S, Boncoraglio GB, Brown RD, Jr, Butterworth AS, Carrera C, Carty CL, Chasman DI, Chen WM, Cole JW, Correa A, Cotlarciuc I, Cruchaga C, Danesh J, de Bakker PIW, Destefano AL, den Hoed M, Duan Q, Engelter ST, Falcone GJ, Gottesman RF, Grewal RP, Gudnason V, Gustafsson S, Haessler J, Harris TB, Hassan A, Havulinna AS, Heckbert SR, Holliday EG, Howard G, Hsu FC, Hyacinth HI, Ikram MA, Ingelsson E, Irvin MR, Jian X, Jimenez-Conde J, Johnson JA, Jukema JW, Kanai M, Keene KL, Kissela BM, Kleindorfer DO, Kooperberg C, Kubo M, Lange LA, Langefeld CD, Langenberg C, Launer LJ, Lee JM, Lemmens R, Leys D, Lewis CM, Lin WY, Lindgren AG, Lorentzen E, Magnusson PK, Maguire J, Manichaikul A, Mcardle PF, Meschia JF, Mitchell BD, Mosley TH, Nalls MA, Ninomiya T, O'Donnell MJ, Psaty BM, Pulit SL, Rannikmae K, Reiner AP, Rexrode KM, Rice K, Rich SS, Ridker PM, Rost NS, Rothwell PM, Rotter JI, Rundek T, Sacco RL, Sakaue S, Sale MM: **Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes.** Nature genetics 2018, **50**(4):524–537.

43. Rabionet Janssen R: **Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes.** Nature genetics 2019,

44. Li H, Khor C, Fan J, Lv J, Yu C, Guo Y, Bian Z, Yang L, Millwood IY, Walters RG, Chen Y, Yuan J, Yang Y, Hu C, Chen J, Chen Z, Koh W, Huang T, Li L: **Genetic risk, adherence to a healthy lifestyle, and type 2 diabetes risk among 550,000 Chinese adults: results from 2 independent Asian cohorts.** The American journal of clinical nutrition 2020, **111**(3):698–707.

45. Myserlis EP, Georgakis MK, Demel SL, Sekar P, Chung J, Malik R, Hyacinth HI, Comeau ME, Falcone GJ, Langefeld CD, Rosand J, Woo D, Anderson CD: **A Genomic Risk Score Identifies Individuals at High Risk for Intracerebral Hemorrhage.** Stroke 2023, **54**(4):973–982.

46. Abraham G, Malik R, Yonova-Doing E, Salim A, Wang T, Danesh J, Butterworth AS, Howson JMM, Inouye M, Dichgans M: **Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke.** Nature Communications 2019, **10**(1):5819.

47. Rutten-Jacobs LC, Larsson SC, Malik R, Rannikmäe K, Sudlow CL, Dichgans M, Markus HS, Traylor M: **Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306 473 UK Biobank participants.** BMJ 2018, **363**:k4168–k4168.

48. Yang S, Sun Z, Sun D, Yu C, Guo Y, Sun D, Pang Y, Pei P, Yang L, Millwood IY, Walters RG, Chen Y, Du H, Lu Y, Burgess S, Avery D, Clarke R, Chen J, Chen Z, Li L, Lv J: **Associations of polygenic risk scores with risks of stroke and its subtypes in Chinese.** Stroke and vascular neurology 2024, **9**(4):399–406.

49. Verbaas C, Fornage M, Bis JC, Choi SH, Psaty BM, Meigs JB, Rao M, Nalls M, Fontes JD, O'Donnell CJ, Kathiresan S, Ehret GB, Fox CS, Malik R, Dichgans M, Schmidt H, Lahti J, Heckbert SR, Lumley T, Rice K, Rotter JI, Taylor KD, Folsom AR, Boerwinkle E, Rosamond WD, Shahar E, Gottesman

RF, Koudstaal P, Amin N, Wieberdink R, Dehghan A, Hofman B, Uitterlinden A, DeStefano AL, Debette S, Xue LT, Beiser A, Wolf PA, DeCarli C, Ikram A, Seshai S, Mosley TH, Longstreth WT, Duijn C, Launer LJ: **Predicting Stroke Through Genetic Risk Functions The CHARGE Risk Score Project.** Stroke (1970) 2014, **45**(2):403–412.

50. HUNT All-In Stroke, CADISP group, International Consortium for Blood Pressure, International Headache Genetics Consortium, International Stroke Genetics Consortium (ISGC) Intracranial Aneurysm Working Group: **Genetic Risk Score for Intracranial Aneurysms: Prediction of Subarachnoid Hemorrhage and Role in Clinical Heterogeneity.** 2023, .

51. MALIK R, BEVAN S, DE STEFANO AL, FORNAGE M, PSATY BM, IKRAM MA, LAUNER LJ, VAN DUIJN CM, SHARMA P, MITCHELL BD, ROSAND J, MESCHIA JF, NALLS MA, LEVI C, ROTHWELL PM, SUDLOW C, MARKUS HS, SESHADRI S, DICHGANS M, HOLLIDAY EG, DEVAN WJ, CHENG Y, IBRAHIM-VERBAAS CA, VERHAAREN BFJ, BIS JC, JOON AY: **Multilocus Genetic Risk Score Associates With Ischemic Stroke in Case–Control and Prospective Cohort Studies.** Stroke 2014, **45**(2):394–402.

52. Hachiya T, Kamatani Y, Takahashi A, Hata J, Furukawa R, Shiwa Y, Yamaji T, Hara M, Tanno K, Ohmomo H, Ono K, Takashima N, Matsuda K, Wakai K, Sawada N, Iwasaki M, Yamagishi K, Ago T, Ninomiya T, Fukushima A, Hozawa A, Minegishi N, Satoh M, Endo R, Sasaki M, Sakata K, Kobayashi S, Ogasawara K, Nakamura M, Hitomi J, Kita Y, Tanaka K, Iso H, Kitazono T, Kubo M, Tanaka H, Tsugane S, Kiyohara Y, Yamamoto M, Sobue K, Shimizu A: **Genetic Predisposition to Ischemic Stroke: A Polygenic Risk Score.** Stroke (1970) 2017, **48**(2):253–258.

53. Hachiya T, Hata J, Hirakawa Y, Yoshida D, Furuta Y, Kitazono T, Shimizu A, Ninomiya T: **Genome-Wide Polygenic Score and the Risk of Ischemic Stroke in a Prospective Cohort: The Hisayama Study.** Stroke (1970) 2020, **51**(3):759–765.

54. Chukwueloka Hezekiah, Blakemore A, Bailey D, Raha Pazoki: **Physical activity reduces the effect of adiposity genetic liability on hypertension risk in the UK Biobank cohort.** *MedRxiv* 2023, **:**.

55. Pazoki R, Dehghan A, Evangelou E, Warren H, Gao H, Caulfield M, Elliott P, Tzoulaki I: **Genetic Predisposition to High Blood Pressure and Lifestyle Factors: Associations With Midlife Blood Pressure Levels and Cardiovascular Events.** Circulation (New York, N.Y.) 2018, **137**(7):653–661.

56. Abraham G, Rutten-Jacobs L, Inouye M: **Risk Prediction Using Polygenic Risk Scores for Prevention of Stroke and Other Cardiovascular Diseases.** Stroke 2021, **52**(9):2983–2991.

57. Phil Beales, Richard H. Scott, Kerrin S. Small, Rohan Taylor, Ana M. Valdes, ChangJiang Xu, Richard Anney, Martin Bobrow, Matthew E. Hurles, Karen Kennedy, Cheryl K. Ridout, Louise Gallagher, Sarah Edkins, Julia Keogh, David K. Jackson, Shane McCarthy, Xueqin Guo, Miriam Schmidts, Paul Flicek, Jun Wang, Karola Rehnström, Jieqin Liang, Juan Pablo Casas, Aaron Day-Williams, Hugh Gurling, Ian N. M. Day, Anja Kolb-Kokocinski, David St Clair, Gaëlle Marenne, Ioanna Tachmazidou, Ewan Birney, James Morris, Robert K. Semple, Parthiban Vijayarangakannan, Elizabeth Stevens, Olli Pietilainen, Steve E. Humphries, Cordelia Langford, Sebahattin Cirak, Xiaosen Guo, Peter McGuffin, David M. Evans, Peter Holmans, Jaana Suvisaari, Martin D. Tobin, Anette Varbo, Ruth Charlton, Michael A. Quail, Petr Danecek, Luis R. Lopes, Stephan Schiffels, Yuanping Du, Alexandros

Onoufriadis, Michael J. Owen, Jon Johnson, Andrew G. McKechanie, Andrew McQuillin, Karim Oualkacha, Guangbiao Wang, Peter Whincup, Klaudia Walter, Gail Clement, So-Youn Shin, Giovanni Gambaro, Jianping Sun, Alireza Moayyeri, Pirro Hysi, Shoumo Bhattacharya, Sally I. Sharp, Gerom: **Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.** Nat Commun 2015, **6**(1):8111.

58. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J: **The Uk biobank resource with deep phenotyping and genomic data.** Nature 2018, **562**(7726):203–209.

59. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'connell J, Cortes A, Welsh S, Mcvean G, Leslie S, Donnelly P, Marchini J: **Genome-wide genetic data on ~500,000 UK Biobank participants.** 2017, **:**.

60. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R: **UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.** PLoS medicine 2015, **12**(3):e1001779.

61. Wtchg (, Oxford), Delaneau O: **UK Biobank Phasing and Imputation Documentation Contributors to UK Biobank Phasing and Imputation.** *Claire Bycroft* 2015, **:**.

62. SHERRY ST: **dbSNP : the NCBI database of genetic variation.** Nucleic Acids Res 2001, **29**:308–311.

63. Flack JM, Adekola B: **Blood pressure and the new ACC/AHA hypertension guidelines.** Trends in Cardiovascular Medicine 2020, **30**(3):160–164.

64. Igo RP, Kinzy TG, Cooke Bailey JN: **Genetic Risk Scores.** Current Protocols in Human Genetics 2019, **104**(1):e95–n/a.

65. Ali A, Jayaraman R, Azar E, Maalouf M: **A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework.** Building and environment 2024, **252**:111268.

66. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V: **Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment.** Medicina (Kaunas, Lithuania) 2020, **56**(9):455.

67. Sperandei S: **Understanding logistic regression analysis.** Biochemia medica 2014, **24**(1):12–18.

68. Bertolaccini L, Pardolesi A, Davoli F, Solli P: **Nanos gigantium humeris insidentes: the awarded Cox proportional hazards model.** Journal of thoracic disease 2016, **8**(11):3464–3465.

69. Austin PC, Giardiello D: **The Impact of Violation of the Proportional Hazards Assumption on the Calibration of the Cox Proportional Hazards Model.** Statistics in medicine 2025, **44**(13-14):e70161–n/a.

70. Sloma M, Syed FJ, Nemati M, Xu KS: **Empirical Comparison of Continuous and Discrete-time Representations for Survival Prediction.** Proceedings of machine learning research 2021, **146**:118–131.

71. Suresh K, Severn C, Ghosh D: **Survival prediction models: an introduction to discrete-time modeling.** BMC medical research methodology 2022, **22**(1):1–18.

Paper 1

3 Chapter 3: Using machine learning to evaluate the value of genetic liabilities in the classification of hypertension within the UK Biobank

3.1 Introduction to Paper 1

This paper was published in the Journal of Clinical Medicine. It involved over 200,000 European participants. The Genetic liabilities were generated from summary statistics and the genetic variants associated with CVD risk factors obtained from genome-wide association studies (GWAS). Various combinations of machine learning models before and after feature selection were tested in order to identify the best classification model for hypertension classification. The models were evaluated using area under the curve (AUC), calibration, and net reclassification improvement in the testing set.

Journal of
Clinical Medicine

MDPI

Article

Using Machine Learning to Evaluate the Value of Genetic Liabilities in the Classification of Hypertension within the UK Biobank

Gideon MacCarthy [1] and Raha Pazoki [1,2,*]

[1] Cardiovascular and Metabolic Research Group, Division of Biomedical Sciences, Department of Life Sciences, College of Health, Medicine and Life Sciences, Brunel University London, London UB8 3PH, UK
[2] MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, St Mary's Campus, Norfolk Place, Imperial College London, London W2 1PG, UK
* Correspondence: raha.pazoki@brunel.ac.uk

Abstract: Background and Objective: Hypertension increases the risk of cardiovascular diseases (CVD) such as stroke, heart attack, heart failure, and kidney disease, contributing to global disease burden and premature mortality. Previous studies have utilized statistical and machine learning techniques to develop hypertension prediction models. Only a few have included genetic liabilities and evaluated their predictive values. This study aimed to develop an effective hypertension classification model and investigate the potential influence of genetic liability for multiple risk factors linked to CVD on hypertension risk using the random forest and the neural network. **Materials and Methods:** The study involved 244,718 European participants, who were divided into training and testing sets. Genetic liabilities were constructed using genetic variants associated with CVD risk factors obtained from genome-wide association studies (GWAS). Various combinations of machine learning models before and after feature selection were tested to develop the best classification model. The models were evaluated using area under the curve (AUC), calibration, and net reclassification improvement in the testing set. **Results:** The models without genetic liabilities achieved AUCs of 0.70 and 0.72 using the random forest and the neural network methods, respectively. Adding genetic liabilities improved the AUC for the random forest but not for the neural network. The best classification model was achieved when feature selection and classification were performed using random forest (AUC = 0.71, Spiegelhalter z score = 0.10, *p*-value = 0.92, calibration slope = 0.99). This model included genetic liabilities for total cholesterol and low-density lipoprotein (LDL). **Conclusions:** The study highlighted that incorporating genetic liabilities for lipids in a machine learning model may provide incremental value for hypertension classification beyond baseline characteristics.

Keywords: the receiver operation characteristic (ROC); area under the curve (AUC)

Citation: MacCarthy, G.; Pazoki, R. Using Machine Learning to Evaluate the Value of Genetic Liabilities in the Classification of Hypertension within the UK Biobank. *J. Clin. Med.* **2024**, *13*, 2955. https://doi.org/10.3390/jcm13102955

Academic Editor: Andrea Dell'Amore

Received: 18 March 2024
Revised: 1 May 2024
Accepted: 7 May 2024
Published: 17 May 2024

1. Introduction

Approximately 1.28 billion people aged 30 to 79 have hypertension worldwide [1], and it continues to rise globally, causing a significant socioeconomic burden due to low awareness and poor control [2]. Hypertension significantly increases the risk of cardiovascular diseases (CVD), including stroke, heart attack, heart failure, and kidney disease, contributing to the global disease burden and premature mortality [1,3,4].

Every year the burden of hypertension and related CVD is increasing in the United Kingdom (UK). As of 2017, hypertension prevalence in England was estimated at around 26.2% among adults [5]. It is responsible for more than half of all strokes and heart attacks, costing the National Health Service (NHS) more than £2.1 billion per year [6].

The current guidelines [7–9] suggest lifestyle modification and the use of blood pressure-lowering medication to prevent hypertension and its consequences. Medication is often successful in lowering blood pressure and reducing the risk of hypertension-related

CVD and stroke. Lifestyle modifications also offer benefits including reduced drug costs, improved control of other comorbidities such as diabetes and hypercholesterolemia, and avoiding preventable pharmacological therapy [10]. The current guidelines have remained silent on the genetic components of hypertension, which are quantifiable at birth and may be used to determine an individual's lifelong disease risk before clinical risk factors are established [11], allowing adequate time to determine lifetime measures to lower hypertension risk, particularly in a high-risk group.

Genome-wide association studies (GWAS) have identified numerous multiple single nucleotide polymorphisms (SNPs) associated with hypertension and/or high blood pressure levels [12–17]. Developing methods to incorporate genetic factors into classification models of hypertension has the potential to improve hypertension classification, management, and control.

Previous studies have used standard statistical techniques or machine learning to predict hypertension [18–29]. However, most of these studies only focused on non-genetic risk factors to predict hypertension [18–23]. The studies that included genetic risk factors only focused on single SNPs at a time [26,30], or gene expression [27], or a single genetic risk score [25]. Our study offers the incorporation of multiple genetic liabilities into machine learning methods above and beyond previous studies. An example of previous studies includes a recent study in rural Chinese populations [25] that incorporated a single hypertension polygenic risk score (PRS) and showed improvement in incident hypertension prediction using several machine learning techniques.

Several studies [31–33] have used a method called multi-polygenic score or meta genetic risk score (metaGRS) that combines several PRSs into regression models for complex diseases, including CVD, and have shown that including multiple genetic factors improves the prediction model's accuracy compared to using one genetic liability. However, these [31–33] did not incorporate the metaGRS within machine learning methods and did not consider hypertension as the outcome in their models. The conventional statistical techniques utilized in previous studies encounter challenges in identifying complex, nonlinear relationships within datasets and exhibit a limited ability to generalize to unseen data (test data). Their constraints emphasize the need for supplementary methodologies, such as machine learning techniques, that offer greater flexibility and robustness in handling complex data structures and achieving accurate predictions. Furthermore, previous studies have not considered the inclusion of genetic liabilities for multiple CVD risk factors to predict hypertension. It is yet to be determined whether machine learning techniques could be applied to enhance hypertension prediction models derived from multiple genetic liabilities for CVD risk factors.

Recent studies have provided evidence for genetic correlations between hypertension and type 2 diabetes [34], adiposity traits [35], lipids traits [36,37], and smoking traits [38]. In the current study, we created genetic liabilities using these risk factors and used machine learning models to evaluate the best combination of genetic liabilities and clinical factors that could optimize the classification of hypertension in the European ancestry population.

2. Material and Method

2.1. Ethical Approval

The UK Biobank (UKB) received ethical approval from the Northwest Multi-centre Research Ethics Committee as a Research Tissue Bank approval, and all the participants provided informed consent. This study is performed using the UKB data under application number 60549. Additionally, we obtained ethics approval from Brunel University London, College of Medicine, and the Life Sciences Research Ethics Committee to work with secondary data from the UKB (reference 27684-LR-Jan/2021-29901-1).

2.2. Study Population

UKB is a prospective observational study with more than half a million participants aged between 40 and 69 years. The participants were recruited between 2006 and 2010

across 22 centres located throughout the United Kingdom (UK). The full description of the UKB study as well as the data collected and a summary of the characteristics are publicly available on the UKB website (www.biobank.ac.uk, accessed on 20 June 2021) and elsewhere by Sudlow and colleagues [39]. In brief, during the recruitment, detailed information about socio-demographics, health status, physician-diagnosed medical conditions, family history, and lifestyle factors was collected via questionnaires and interviews. Several physical measurements, including height, weight, body mass index (BMI), waist–hip ratio (WHR), systolic blood pressure (SBP), and diastolic blood pressure (DBP), were obtained. The records of participants in the UKB project were accordingly linked to Hospital Episode Statistics (HES) data, as well as national death and cancer registries.

The current study is based on a subset of unrelated individuals of European ancestry (n = 244,718; Figure 1). In brief, we used 40 genetic principal components created centrally by the UKB and applied the k-means clustering method on 502,219 UKB participants to identify individuals of European descent. We then obtained genetic data from the individuals who had passed the UKB internal quality control and had genotype data (n = 459,042). We excluded individuals (n = 25,340) who had been diagnosed with a stroke, heart attack, or angina before or at baseline. This strategy helps to adjust for pre-existing CVDs and minimizes the possibility of confounding. We excluded participants who had withdrawn their consent (n = 61), pregnant individuals, or those uncertain about their pregnancy status (n = 278). We additionally excluded individuals with mismatched genetics and self-reported (n = 320) to avoid potential inconsistencies in data reporting. Using the kinship cut of 0.0884 for third-degree relatives, we further excluded participants who were up to second-degree related (n = 33,369). Furthermore, individuals who were on cholesterol-lowering medication (n = 34,243), stopped smoking or drinking due to health reasons or doctor's advice (n = 58,752), and participants with missing data on the potential confounders (n = 61,961) were excluded from the dataset, leaving a final 244,718 unrelated individuals of European ancestry for our analyses.

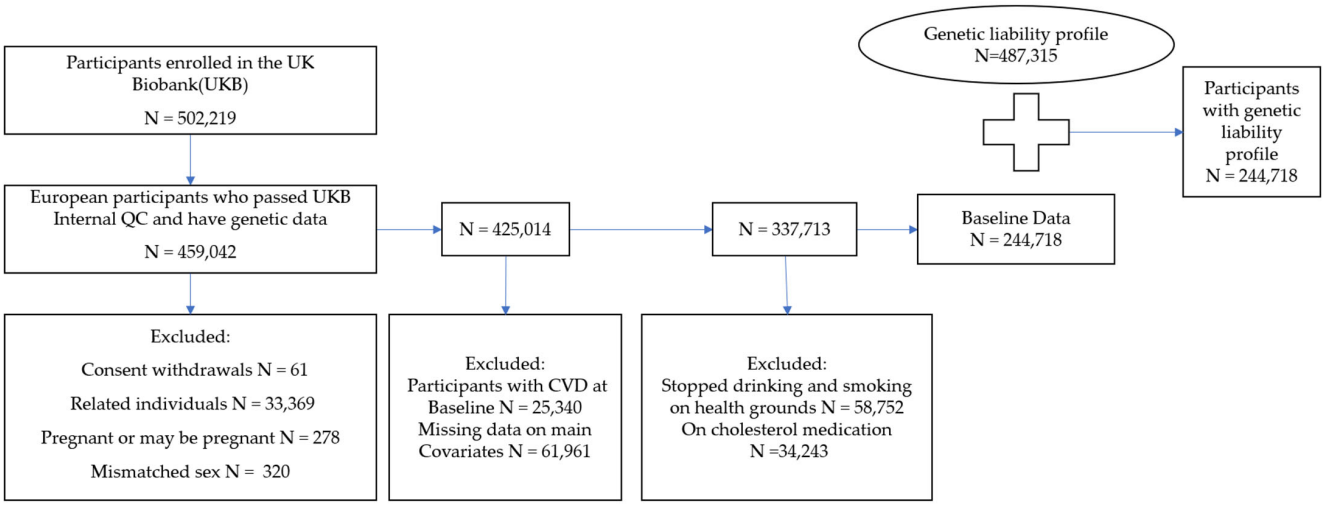


Figure 1. The flowchart of the study participant selection. UK Biobank (UKB) data had 502,219 participants at the beginning of this study. We extracted 459,042 participants of European ancestry who have passed UKB internal quality control (QC) and have genetic data. The final dataset included 244,718 participants who met the inclusion criteria for whom a genetic liability profile was generated.

2.3. Genotyping and Imputation

The UKB conducted all the DNA extraction, genotyping, and imputation. The detailed processes have been discussed elsewhere [40–42]. In brief, blood samples from participants were obtained at UKB assessment centres, and DNA was extracted and genotyped using the UKB Axiom Array. The genotype imputation was conducted by UKB using the IMPUTE4 tool. Three reference panels—Haplotype Reference Consortium, UK10K, and 1000 Genomes

phase 3—were used for the imputation. The genetic principal components and kinship coefficients were calculated centrally by UKB to account for population stratification and identify related individuals [40,42].

2.4. Definition of the Outcome

Our main outcome is hypertension, which was defined as (1) the presence of a recorded SBP $\geq$ 140 mmHg or a DBP $\geq$ 90, or (2) hypertension diagnosed by a doctor, or (3) a record of using blood pressure-lowering medication at baseline [43]. In the UKB, two blood pressure readings were obtained a few minutes apart using a standard automated device or manual sphygmomanometer (www.ukbiobank.ac.uk, accessed 20 June 2021). We calculated both mean SBP and mean DBP from two automated or two manual readings of blood pressure measurements. For participants with one manual and one automated blood pressure reading, the average of these two values was used. For individuals with a single blood pressure measurement (one manual or one automated blood pressure reading), the single measurement was used for approximating the participant's blood pressure value. For the participants who self-reported to be taking blood pressure-lowering medication, we added 15 mmHg to SBP and 10 mmHg to DBP [44]. The participants with missing blood pressure readings were excluded.

2.5. Demographics and Clinical and Lifestyle Features

In the statistical analysis, factors such as age, sex, BMI, diabetes mellitus, total cholesterol (TC), low-density lipoprotein (LDL), high-density lipoprotein (HDL), smoking status, drinking status, and sedentary lifestyle were included. Diabetes was defined as a record of diabetes diagnosed by a doctor, or using insulin medication, or a record of serum level of haemoglobin A1c (HbA1c) $\geq$ 48 mmol/mol (6.5%), or glucose level $\geq$ 7.0 mmol/dL [45]. Smoking and alcohol consumption data were collected through a self-reported questionnaire by the UKB and were classified into current, previous, and never.

We calculated a sedentary lifestyle variable by approximating the total self-reported hours per day the participants spent on (1) driving, (2) using a computer, and (3) watching television. We considered 30 min of sedentary behavior if individuals indicated that they spent less than an hour per day driving, or watching television, or using a computer. In this study, the demographics and clinical and lifestyle (non-genetic) features were selected as conventional risk factors for CVD that have been used often in previously published work of ours and others [28,44].

2.6. Computation of Genetic Liabilities

SNP Selection

We selected a list of genetic variants in the form of SNPs (Table 1 and Supplementary Data S1–S10) that were previously identified as associated with ten CVD risk factors, including type 2 diabetes [46], two adiposity traits [47,48], three smoking traits [49], and four lipid traits [50] at a GWAS significant threshold (p-value $< 5.0 \times 10^{-8}$) in the European population. Linkage disequilibrium (LD) measures the non-random linkage of alleles at different loci on the same chromosome in a population. SNPs are said to be in LD when the frequency of association between their alleles exceeds what would be expected from a random assortment [51]. LD between two loci is determined statistically using metrics like r^2. This metric measures the level of connection between alleles at the two loci. The SNPs used in calculating the genetic liabilities were pruned with the LD pruning procedure employed in SNPclip incorporated within the LDlink online tool (https://ldlink.nih.gov/?tab=home, accessed on 20 July 2021). A minor allele frequency (MAF) of 0.01 and r^2 threshold of 0.1 was used in LD pruning. Duplicate SNPs, not biallelic SNPs, SNPs with MAF less than 0.01, and SNPs in LD with other SNPs ($r^2 > 0.1$) were excluded. We used the final list of selected LD-pruned SNPs to estimate genetic liabilities for all the ten traits in the current study using PLINK version 1.9 [52]. To allocate weight to each SNP, we used the effect sizes estimated for the association of the SNPs with each of the traits mentioned

in Table 1. The effect sizes (Supplementary Data S1–S10) were obtained from previously published, publicly available GWAS summary statistics data provided for these SNPs within the GWAS Catalog website (https://www.ebi.ac.uk/gwas/, accessed on 12 July 2021). PLINK uses a weighted method, where the effect size (beta coefficient) of each SNP is considered as weight and is multiplied by the number of risk alleles an individual carries. The product is then summed across all SNPs to produce genetic liability for each person. We standardized all the genetic liabilities (mean-centred with standard deviation 1).

Table 1. Published GWAS SNPs for calculating the genetic liabilities.

Trait Category	Genetic Liability	Study (Publication Year)	Number of SNPs	Reference
Smoking	Smoking initiation	Liu et al., 2019 [49]	311	Liu et al., 2019 [49]
	Smoking cessation	Liu et al., 2019 [49]	16	Liu et al., 2019 [49]
	Smoking heaviness	Liu et al., 2019 [49]	38	Liu et al., 2019 [49]
Diabetes	Type 2 diabetes	Mahajan et al., 2018 [46]	210	Mahajan et al., 2018 [46]
Adiposity	BMI	Winkler et al., 2016 [47]	159	Winkler et al., 2016 [47]
	WHR	Shungin et al., 2015 [48]	39	Shungin et al., 2015 [48]
Lipid traits	TC	Surakka et al., 2015 [50]	36	Surakka et al., 2015 [50]
	HDL	Surakka et al., 2015 [50]	19	Surakka et al., 2015 [50]
	LDL	Surakka et al., 2015 [50]	30	Surakka et al., 2015 [50]
	Triglycerides	Surakka et al., 2015 [50]	25	Surakka et al., 2015 [50]

GWAS: genome-wide association studies, SNPs: single nucleotide polymorphisms, BMI: body mass index, WHR: waist–hip-ratio, TC: total cholesterol, HDL: high-density lipoprotein, LDL: low-density lipoprotein.

2.7. Statistical Analysis

We summarized the categorical variables using frequencies and percentages, and the continuous variables were expressed as the mean (SD). When comparing the characteristics differences between the hypertensive and non-hypertensive groups, the nonparametric test (Wilcoxon rank sum test) was utilized for continuous variables as the assumptions for the parametric *t-test* may not be met. The chi-squared test was applied to compare the hypertensive and non-hypertensive groups for categorical variables. We used univariable logistic regression to determine the strength and direction of the relationship between individual features with hypertension without considering other variables. We also used multivariable logistic regression to assess the independent effects or association of each feature while controlling for the effects of other features in the model. The statistical significance of the association was defined, where the associations demonstrated a 2-sided *p*-value less than 0.05 (see Supplementary Tables S1 and S2).

2.8. Data Preprocessing and Splitting

We excluded participants who had missing values for essential variables. Features included in our machine learning algorithm are presented in Table 2.

All categorical variables were labelled, including gender (0 = female, 1 = male), smoking status (0 = never, 1 = previous, 2 = current), alcohol consumption status (0 = never, 1 = previous, 2 = current), diabetes (0 = no, 1 = yes), and our outcome variable, hypertension (0 = no, 1 = yes). Our numerical variables were all measured on different scales. To ensure that all the numerical variables contribute equally to our model [53], we scaled them to a given range, using a "min–max" approach.

In machine learning, data splitting is a common practice used for evaluating the performance of a prediction model. This involved splitting the available dataset into training and testing sets. The training set is for training the machine learning model, and the testing is used to assess the model's performance. In this study, we employed the train-test split approach [54] to randomly partition the dataset at a ratio of 70:30 (Figure 2)

into a training set (70%; n = 171,304; case = 81,967 and control = 89,337) and a testing set (30%; n = 73,414) using the "*createDataPartition*" function in the R-package version 4.2.2. This approach ensures both our training set and testing set capture the underlying distribution of the data.

Table 2. Features included in the machine learning algorithm.

Feature Category	Feature Type	Feature
Characteristics features		• Sex • Age
Lifestyle-related features	Phenotype	• Smoking status • Sedentary lifestyle • Drinking status
	Genetic	• Genetic liability for smoking heaviness • Genetic liability for smoking cessation • Genetic liability for smoking initiation
Diabetes-related features	Phenotype	• Diabetes (see methods for definition)
	Genetic	• Genetic liability for type 2 diabetes
Adiposity-related features	Phenotype	• BMI
	Genetic	• Genetic liability for BMI • Genetic liability for WHR
Lipid-related features	Phenotype	• TCL • DLH • DL
	Genetics	• Genetic liability for LDL • Genetic liability for HDL • Genetic liability for TC • Genetic liability for triglycerides

BMI: body mass index, WHR: waist–hip ratio, TC: total cholesterol, HDL: high-density lipoprotein, LDL: low-density lipoprotein.

The models were trained in the training set, and the performance of the models in terms of discrimination ability (defined as the model's capacity to distinguish between persons with and without outcomes) was assessed in the testing set (n = 73,414; Figure 2). To this end, we constructed the receiver operating characteristic curve (ROC) for each model and calculated the area under the curve (AUC) with 95% confidence intervals (CIs) [55–57]. The AUC ranges from 0.5 to 1.0, with 0.5 indicating no better discrimination than chance and 1.0 representing perfect discrimination power.

2.9. Handling Data Imbalance

Models trained on imbalanced datasets may become biased towards the dominant class, predicting the minority class incorrectly [58]. A binary classifier, which is the case in the current study, trained on a balanced dataset typically outperforms a model trained on an imbalanced dataset [59]. An imbalanced training dataset may lead to overfitting the majority class due to their higher prior probability [60]. This means that the minority class may be misclassified more frequently as compared with the majority class. This issue could lead to incorrect prediction and that some model performance metrics, such as accuracy, may be distorting the conclusions [60]. Balancing the training set is an approach that could prevent overfitting, reduce bias, and ensure that the model learns to successfully classify groups in addition to improving the accuracy of prediction on the testing data. To balance the number of events in the training set before training the models, we utilized the random over-sampling using a bootstrapping method implemented within the "*ROSE*" (Random Over-Sampling Examples) package [61]. To deal with imbalanced data and balance the class distribution in the dataset, the ROSE package generates synthetic samples for the minority class [61].

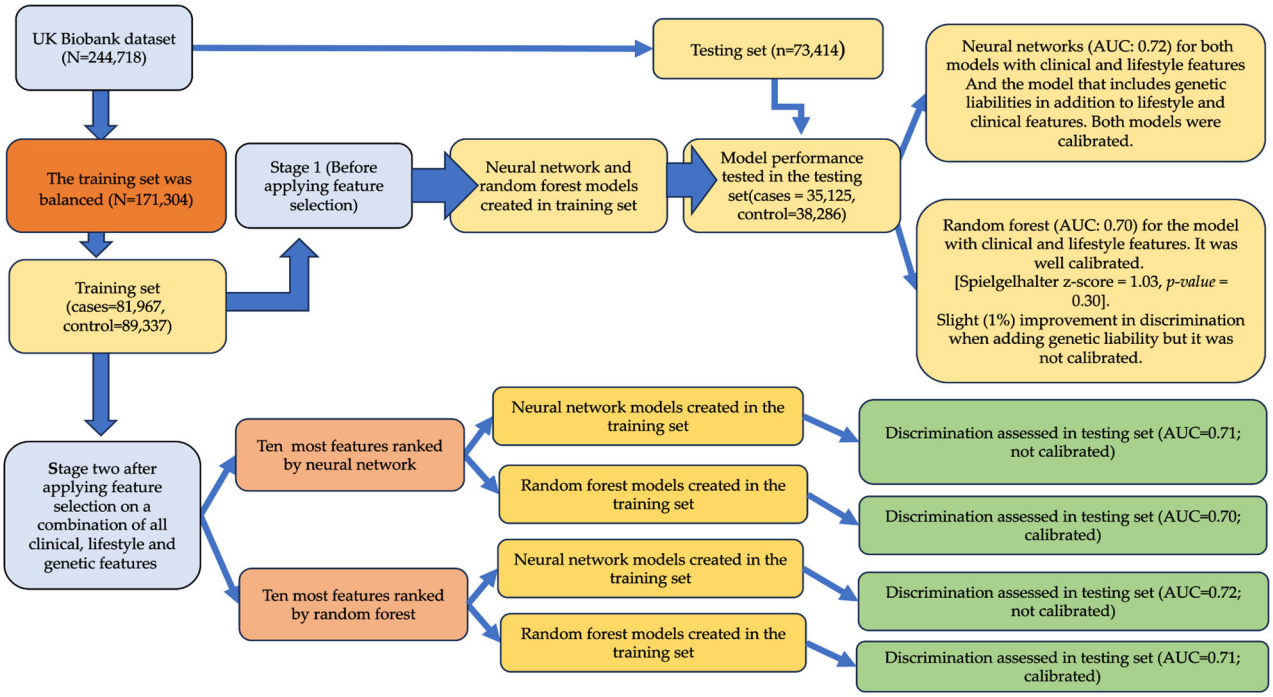


Figure 2. Overview of the study and the construction of machine learning models: The flowchart of the study design. The data were split into training and testing sets. Stage one models were built in the training set without selection, and their performances were assessed in the testing set. Stage two models were built after the feature selection technique was applied, and their performances were evaluated in the testing set.

2.10. Machine Learning Model Construction

In comparison with traditional statistical techniques, machine learning algorithms are flexible and free of prior assumptions (e.g., the type of error distribution) and can capture the complicated, nonlinear relationships between predictors. These algorithms automate decision-making processes using models that have been trained on historical data [62]. They can analyse various data types and integrate them into predictions for disease risk [63]. Machine learning algorithms with and without genetic data were used in the prediction of hypertension in the European population, including support vector machines, decision trees, random forest, neural network, and extreme gradient boosting (XGBoost) [64]. For this study, we considered two machine learning-based classifiers, the random forest and the neural network, which have been shown in many studies to have been most promising in the classification of hypertension among individuals of European ancestry [64].

The random forest is a powerful machine learning algorithm that constructs an ensemble or forest of decision trees that are often trained using the bagging method. Each decision tree is constructed on a random subset of the training set and a random subset of the features. This keeps the trees from becoming overly correlated and, hence, overfitting the data [65]. Using the training set, the random forest models were constructed with the "*ranger*" package in the R-programme [66] with hyperparameters set to 500 trees and 10 nodes. The optimal model was selected based on an out-of-bag (OOB) estimate of the error rates in the training set. In a random forest model, the maximum number of features that can be considered for splitting at each node of the decision trees within the ensemble was determined and reported using the "*mtry*" parameter within the "*ranger*" package (Supplementary Table S1).

The neural network is another powerful machine learning algorithm that automatically learns from patterns between the inputs and the output within the data [67]. The neural network consists of interconnected processing nodes organized in three layers: input,

hidden, and output layers (Supplementary Figure S1). The input layer is connected to the hidden layer with updated weight, which is then connected to the output layer [68]. In the construction of our neural network models, the optimal number of hidden layers was identified as 5 hidden layers. The neural network classifiers were constructed with the "nnet" function [69] using 5-fold cross-validation implemented in the R-programme "caret" package. We used the following hyperparameters: The regularization parameter used to prevent overfitting in the neural network by penalizing large weights (decay) was set to 0. Setting it to 0 means that no weight decay is applied. The maximum number of iterations (epochs) for training the neural network (maxit) was set to 100, and the maximum number of weights in the neural network (maxNWTs) was set to 1000.

Overfitting and underfitting are two frequent machine learning issues that can have a significant influence on model performance and generalizability [70]. Overfitting occurs when a model fits the training set but performs poorly on the testing or an unknown dataset, resulting in low training error but high test error. To minimize overfitting within the neural network models and to produce more reliable estimations of the models' prediction abilities on the testing set [71,72], we performed a 5-fold cross-validation on the balanced training set (n = 171,304). The ROC [73] was used to select the optimal model (based on the largest ROC value estimated at 0.70; see Supplementary Table S2). This implies that models with ROC values greater than 0.70 in the testing set would improve prediction.

We adopted a two-stage approach in the construction of the machine learning algorithm.

2.10.1. Stage One Models

In stage one, we built models without feature selection in the training set (n = 171,304). In the first subset of models, we used the random forest method [65,66], and in the second subset of models, we used the neural network (Figure 2). For each of these methods, we used two different sets of features: (1) conventional risk factors that included baseline characteristics (age, sex, BMI, diabetes mellitus, smoking status, drinking status, TC, HDL, LDL, and sedentary lifestyle; and (2) full set of features included all the baseline characteristics above together with additional genetic variables, including ten genetic liabilities.

The optimal number of features used for splitting at each of the decision trees was identified as three features in the construction of the random forest model with ten conventional risk factors vs. four features in the random forest model with the additional ten genetic liabilities. Both random forest models above showed a prediction error of 0.22.

In the testing set, we evaluated and compared the performance of (1) the random forest model with and without genetic liabilities and (2) the neural network model with and without genetic liabilities. The performance of our machine learning models was evaluated using the AUC, accuracy, sensitivity (recall), and F1 score.

2.10.2. Stage Two Models

In stage two, to improve model performance, we utilized a feature selection strategy to select the most relevant features, eliminate unnecessary noise or random fluctuation in the data, and prevent the problem of overfitting (induced by the presence of irrelevant features). Both the random forest and neural network approaches were used as the feature selection method. The most important features were ranked based on their importance score and illustrated using variable importance plot (*vip*) function within the *caret* package. Regardless of the method used in the feature selection step, we further used the top ten most important features identified to further develop classification models using random forest and neural network. This approach created four different analysis paths to hypertension classification including the path (1) where the feature selection model was random forest, and the classifying method was random forest as well; path (2) where random forest was used as the feature selection method, and the classification method was neural network; path (3) where the feature selection model was neural network, and the classification method was neural network as well; path (4) where neural network was the feature selection method, and the classification method was random forest (see Figure 2). In the testing set, we used

the AUC (see above) to assess the performance of these four models built with the ten most important features selected. Stage one and stage two resulted in the construction and testing of a total of eight models.

2.11. Model Performance Assessment Using Calibration

We used a calibration curve and Spiegelhalter z score test to examine the models' calibration [74,75]. Model calibration measures the ability of a model to accurately predict an outcome [76,77]. In the calibration curve, the *Y*-axis represents the observed probability, and the *X*-axis represents the predicted probability of developing a disease. The calibration curve includes a diagonal line (the ideal line), which is the prediction of the ideal model. A model is said to be well-calibrated if the calibration curve stays close to the line of perfect calibration (45 degrees with an intercept of 0 and a slope of 1). Overestimation and underestimation are identified by a curve below and above the ideal calibration line, respectively. The Spiegelhalter z test is a statistical test used to assess the calibration accuracy of a risk prediction model. A perfectly calibrated model (i.e., when the predicted probabilities match the observed values) has a Spiegelhalter z score of zero, while a value close to zero indicates good calibration, and a value far from zero indicates poor calibration. A positive Spiegelhalter z score indicates that the model is over-calibrated (i.e., the predicted probability of the outcome is too high), while a negative Spiegelhalter z score indicates that the model is under-calibrated (i.e., the predicted probability of the outcome is too low).

To confirm the overall accuracy of the models, we also calculated the Brier score [78], which is the mean square error (MSE) between observed and predicted outcomes. The Brier score evaluates both the calibration and discrimination ability of a model [77]. The scores range from 0 to 1, with lower scores suggesting superior calibration. Brier scores approaching 0 imply that the model has been adequately calibrated and discriminated. We used the validation probability (*val.prob*) function from the "*rms*" package in the R-programme to generate calibration curves, Spiegelhalter z test, and Brier score.

2.12. Net Reclassification Index and Integrated Discrimination Index

We assessed the performance of well-calibrated models using the net reclassification index and integrated discrimination index statistics. The net reclassification improvement is a commonly used metric to compare the relative ability of two models to classify individuals as low- and high-risk [79]. A positive net reclassification index value indicates that the new model correctly reclassifies more individuals into higher- or lower-risk categories compared to the old model. Conversely, a negative net reclassification index value suggests that the old model is better at reclassifying individuals than the new model.

The integrated discrimination index statistic is used to measure the improvement in the ability of two models to distinguish between event and non-event [80,81]. A positive integrated discrimination index value implies an improvement in the model's discriminative ability, while a negative integrated discrimination index value suggests a deterioration in the discriminative ability of the new model. In this study, we used the "*reclassification*" function from the "*PredictABEL*" packages in the R-programme to obtain the net reclassification index and integrated discrimination index values. The discrimination ability, calibration, and reclassification results are depicted further in Figure 2. All the analysis was performed with R-program (www.r-project.org; access data December 2022) version 4.2.2. For reproducibility, we set the seed of the random number generator to a value of 500 throughout this analysis. The code for the analyses has been generated and accessible to the public through Github links below:

1. https://github.com/GMaccarthy/NN_with_Imbalanced_Trainingset, accessed on 30 April 2024;
2. https://github.com/GMaccarthy/NN_with_Balanced_Trainingset, accessed on 30 April 2024;
3. https://github.com/GMaccarthy/RF_Balanced_trainingset, accessed on 30 April 2024;
4. https://github.com/GMaccarthy/RF_Imbalaced_Trainingset, accessed on 30 April 2024.

3. Results

3.1. Baseline Characteristics of the Participants

A total of 244,718 unrelated individuals of European ancestry from the UKB were included in this study (Table 3). The average age of the participants was 55.4 ± 7.98 years old, and 141,931 (58.0%) participants were female. The sample contained 7011 (2.9%) participants with diabetes. The majority (n = 229,539; 93.8%) of the participants reported to be current alcohol drinkers, and 164,847 (67.4%) reported to have never smoked. The average BMI was 26.8 (4.58) kg/m^2. The sample included 117,095 (47.8%) participants with hypertension. There were statistically significant differences in all baseline characteristics between the hypertensive and non-hypertensive groups (Table 3). More women were hypertensive than men (52.4% vs. 47.6%; *p*-value < 0.001). The hypertensive participants were older (57.6 ± 7.53 vs. 53.4 ± 7.84 years; *p*-value < 0.001), had higher BMI (28.0 ± 4.83 kg/m^2 vs. 25.8 ± 4.06 kg/m^2; *p*-value < 0.001), had higher TC levels (6.05 ±1.06 vs. 5.79 ± 1.04 mmol/L; *p*-value < 0.001), and spent more hours per day having sedentary lifestyle (4.87 ± 2.39 vs. 4.50 ± 2.33 h per day; *p*-value < 0.001) than the non-hypertensive participants.

Table 3. Baseline characteristic of the UKB participants within the overall sample and hypertensive subgroups.

	Hypertensive	**Non-Hypertensive**	**Overall**	***p*-Value**
	n = 117,095	n = 127,623	n = 244,718	
Diabetes diagnosed by a doctor:				
YES; N (%)	4697 (4.00%)	2314 (1.80%)	7011 (2.9%)	<0.001
NO; N (%)	112,398 (96.0%)	125,309 (98.2%)	237,707 (97.1%)	
Age (years); mean (SD)	57.6 (7.53)	53.4 (7.84)	55.4 (7.98)	<0.001
BMI (kg/m^2); mean (SD)	28.0 (4.83)	25.8 (4.06)	26.8 (4.58)	<0.001
TC (mmol/L); mean (SD)	6.05 (1.06)	5.79 (1.04)	5.91 (1.06)	<0.001
HDL (mmol/L); mean (SD)	1.46 (0.38)	1.51 (0.38)	1.49 (0.38)	<0.001
LDL (mmol/L); mean (SD)	3.84 (0.81)	3.62 (0.80)	3.73 (0.81)	<0.001
Sedentary lifestyle (h/day); mean (SD)	4.87 (2.39)	4.50 (2.33)	4.68 (2.37)	<0.001
Sex:				
Male; N (%)	55,686 (47.6%)	47,101 (36.9%)	10,2787 (42.0%)	<0.001
Female; N (%)	61,409 (52.4%)	80,522 (63.1%)	141,931 (58.0%)	
Drinking status:				
Current; N (%)	109,655 (93.6%)	119,884 (93.9%)	229,539 (93.8%)	<0.001
Never; N (%)	4052 (3.46%)	3967 (3.11%)	8019 (3.3%)	
Previous; N (%)	3388 (2.89%)	3772 (2.96%)	7160 (2.9%)	
Smoking status:				
Current; N (%)	37,458 (32.0%)	39,476 (30.9%)	76,934 (31.4%)	<0.001
Never; N (%)	78,292 (66.9%)	86,555 (67.8%)	164,847 (67.4%)	
Previous; N (%)	1345 (1.15%)	1592 (1.25%)	2937 (1.2%)	

Table is generated with gtsummary package using Pearson's chi-squared test and Wilcoxon rank sum test. SD: standard deviation, BMI: body mass index, TC: total cholesterol, HDL: high-density lipoprotein, LDL: low-density lipoprotein.

All the demographic, clinical, and lifestyle features included in the study had a statistically significant association with hypertension (Table 3, Supplementary Tables S3 and S4).

3.2. Stage One Models

In stage one (Figure 2), the models incorporating conventional CVD risk factors (i.e., age, sex, BMI, diabetes, smoking status, drinking status, TC, HDL, LDL, and sedentary lifestyle) achieved AUCs of 0.70 (95% CI = 0.70, 0.71; Table 4 and Figure 3), accuracy of 0.65 (95% CI = 0.65, 0.66), a sensitivity of 0.68, and F1-Score of 0.64 using the random forest method (Table 4). The calibrations measured by Spiegelhalter's z score were 1.03 (p-value = 0.30, calibration slope = 0.98) for random forest (Table 4 and Supplementary Figure S2). We observed AUC of 0.72 (95% CI = 0.71, 0.72), accuracy of 0.66 (95% CI = 0.65, 0.66), a sensitivity of 0.69, and F1-Score of 0.66 using neural network (Table 4 and Figure 3). Spiegelhalter's z score was estimated as −14.39 (p-value = 6.4×10^{-47}, calibration slope = 1.18) using neural network (Table 4, Figure S3).

The addition of genetic liabilities resulted in a slight improvement in the AUC only in random forest model (AUC = 0.71; Table 4 and Figure 4). We observed a Spiegelhalter's z score of −5.64 (p-value = 1.7×10^{-8}, calibration slope = 1.06; Table 4 and Supplementary Figure S2). A Spiegelhalter's z score of −14.44 (p-value = 3.0×10^{-47}, calibration slope = 1.18) was observed for neural network models.

Table 4. Discrimination and calibration results of the models applied to the testing set.

Classification Models	Numb of Features	R^2	AUC% (95% Cl)	Brier Score	Spiegel Halter z Score	Spiegel Halter p-Value	Slope	Intercept	Accuracy % (95% Cl)	Sensitivity (Recall)	F1 Score
Models with conventional risk factors											
Random forest	10	0.17	0.70 (0.70, 0.71)	0.22	1.03	0.30 *	0.98	0.04	0.65 (0.64, 0.65)	0.68	0.64
Neural network	10	0.19	0.72 (0.71, 0.7)	0.21	−14.39	6.4×10^{-47}	1.18	0.08	0.66 (0.65, 0.66)	0.69	0.66
Models with conventional risk factors and genetic liabilities											
Random forest	20	0.18	0.71 (0.71, 0.72)	0.22	−5.64	1.7×10^{-8}	1.06	−0.04	0.65 (0.64, 0.65)	0.68	0.65
Neural network	20	0.19	0.72 (0.71, 0.72)	0.21	−14.44	3.0×10^{-47}	1.18	0.07	0.66 (0.65, 0.66)	0.68	0.66
Random forest as feature selection method											
Random forest	10	0.17	0.71 (70, 0.71)	0.22	0.10	0.92	0.99	−0.04	0.65 (0.64, 0.65)	0.66	0.64
Neural network	10	0.18	0.72 (0.71, 0.72)	0.21	−15.51	3.1×10^{-54}	1.20	−0.09	0.66 (0.65, 0.66)	0.69	0.66
Neural network as feature selection method											
Random forest	10	0.16	0.70 (0.70, 0.71)	0.22	−0.44	0.66	1.00	−0.04	0.64 (0.64, 0.65)	0.66	0.64
Neural network	10	0.17	0.71 (0.70, 0.71)	0.22	−13.80	1.6×10^{-43}	1.18	−0.08	0.65 (0.65, 0.66)	0.69	0.65

* p-value > 0.05 (test is not significant) good calibration. Model with conventional risk factors included age, sex, BMI, diabetes, smoking status, drinking status, TC, HDL, LDL, and sedentary lifestyle. Discrimination is measured by the AUC. Accuracy is the percentage of true predictions made by our model out of all predictions made; the F1 is a single score that balances both precision and recall (sensitivity). Recall is the proportion of true positive predictions among all positive instances in the dataset. The Brier score is a combined measure of discrimination and calibration. Calibration is measured by the Spiegelhalter z test, logistic slope, and intercept. BMI: body mass index, AUC: area under the curve, TC: total cholesterol, HDL: high-density lipoprotein, LDL: low-density lipoprotein.

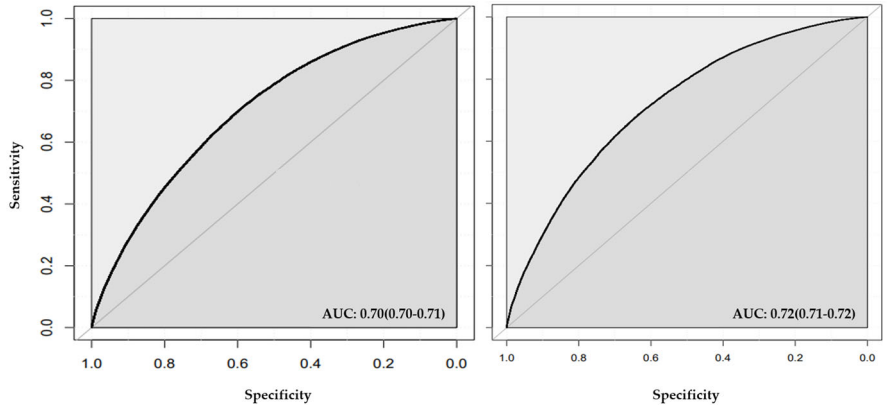


Figure 3. ROC plot for models with conventional risk factors in stage one. The figure shows the area under the curve (AUC) for both random forest (left panel) and neural network (right panel).

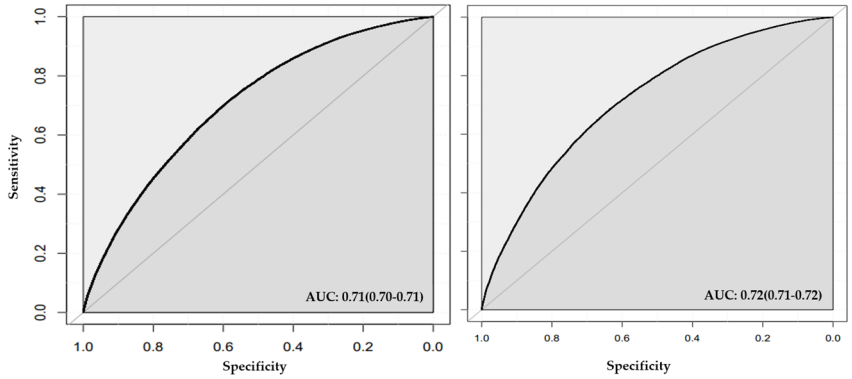


Figure 4. ROC plot for stage one models including conventional risk factors and genetic liabilities. The figure shows the AUC for random forest (left panel) and neural network (right panel).

3.3. Stage Two Models

In stage two (Figures 5 and 6), random forest feature selection identified feature age as the most important classifying feature for hypertension, followed by sex, BMI, TC, LDL, sedentary lifestyle, HDL, TC genetic liability, LDL genetic liability, and smoking status. (Supplementary Figure S4). Feature selection using neural network identified HDL as the most important feature, followed by TC, LDL, sedentary lifestyle, LDL genetic liability, BMI, TC genetic liability, age, WHR genetic liability, and HDL genetic liability (Supplementary Figure S5).

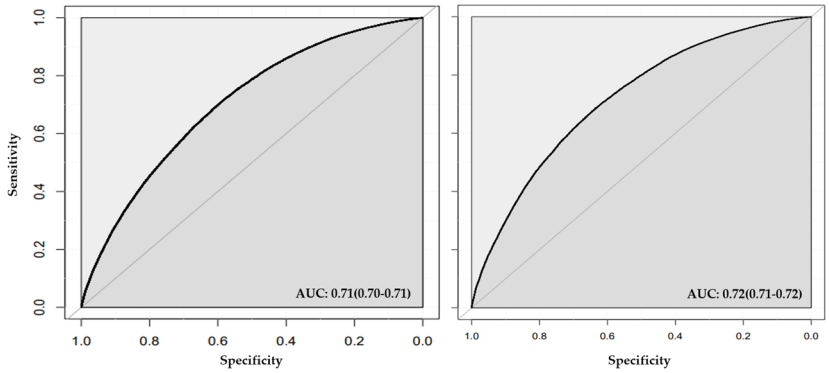


Figure 5. ROC plot for stage two models created with features selected by random forest that included conventional risk factors and genetic liabilities. Area under the curve (AUC) is illustrated for classification by random forest (left panel) and neural network (right panel).

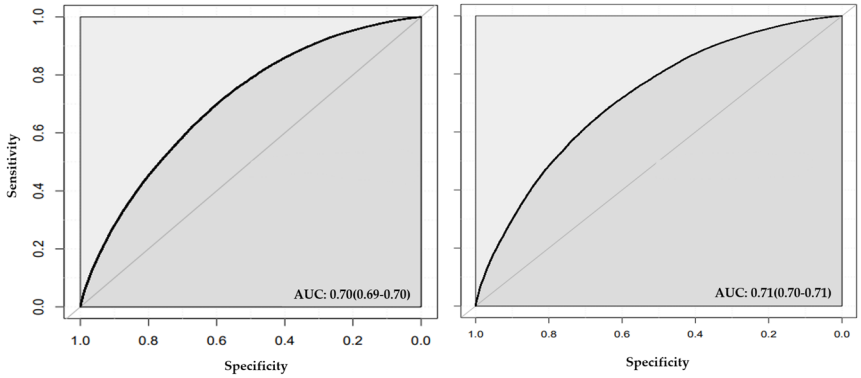


Figure 6. ROC plot for stage two models created with features selected by neural network that included conventional risk factors and genetic liabilities. Area under the curve (AUC) is illustrated for classification by random forest (left panel) and neural network (right panel).

The model in stage two that was well-calibrated and achieved an improved AUC among the four models developed in this stage was the model built with the important features (Supplementary Figure S4) selected and classified using random forest (see Methods). The model achieved an AUC of 0.71 (95% Cl = 0.70, 0.71) and a Spiegelhalter's z score of 0.10 (p-value = 0.92, calibration slope = 0.99; Supplementary Figure S6). The model showed accuracy of 0.65 (95% CI = 0.64, 0.65), sensitivity of 0.66, and F1-Score of 0.64 (Table 4 and Figure 5).

3.4. Reclassification Index Analysis

Three models with a random forest classifier, including one from stage one and two from stage two analysis, were identified as well-calibrated. These models were included in the reclassification index analysis where the model from stage one was used as the reference (that is the model which included all conventional CVD risk factors and used random forest as classifying method; Figure 2).

The stage two model which was built using random forest as both feature selection and classification method (Table 5) showed a slightly improved reclassification compared with the reference model indicated by a net reclassification index of 0.06 (95% CI = 0.05, 0.08; Table 5). This model showed an integrated discrimination index of 1.7×10^{-3} (95% CI = 9.0×10^{-4}, 2.5×10^{-3}; Table 5).

Table 5. Net reclassification and integrated discrimination index.

Feature Selection Method	Classification Method	$NRI^{>0}$ (95% Cl) p-Value	IDI (95% Cl) p-Value
None	* Random forest	Ref	Ref
Random forest	Random forest	0.06 (0.05, 0.08) p-value < 0.00001	1.7×10^{-3} (9.0×10^{-4}, 2.5×10^{-3}) p-value = 1.0×10^{-5}
Neural network	Random forest	−0.10 (−0.12, −0.09) p-value < 0.00001	−0.01 (-9.3×10^{-4}, −0.01) p-value < 0.00001

* The random forest model with all conventional risk factors was selected as a reference model. $NRI^{>0}$: continuous net reclassification index; IDI: integrated discrimination index, Cl: confidence interval, Ref: reference.

The stage two model which was built using neural network as feature selection method and random forest as classification method showed a deteriorated reclassification compared to the reference model, indicated by a net reclassification index of −010 (95%

CI = 0.12, 0.09; Table 5). This model showed an integrated discrimination index of −0.01 (95% CI = -9.3×10^{-4}, −0.01; Table 5).

4. Discussion

This is the first large-scale research that has been conducted on hypertension classification using machine learning that investigates the prediction value of a combination of genetic liabilities for type 2 diabetes, adiposity traits, lipid traits, and smoking traits in a single model. Aided by machine learning, we used a European dataset with 244,718 participants from the UKB and identified the best integrated predictive models for the classification of hypertension. We found that incorporating multiple genetic risk factors into prediction models could lead to a minor but statistically significant improvement in the classification ability and reclassification of the models beyond conventional risk factors. Of all the genetic liabilities we considered, those estimated for TC and LDL cholesterol were identified to be a combination that could improve the classification of hypertension compared with the model without any genetic factors. This is the first study that identifies the predictive value of the genetic liability of lipid traits in the hypertension classification. Several cohort studies have found a link between high cholesterol levels [82,83] as well as dyslipidaemia [84,85] and an increased risk of developing hypertension. Dyslipidaemia is known to impair the functional and structural features of the arteries and cause atherosclerosis [86]. These changes may compromise blood pressure control, predisposing individuals with dyslipidaemia to hypertension.

Previous literature has only described traditional statistical techniques and machine learning models to predict/classify hypertension, mainly using non-genetic risk factors [18–24]. The studies that included genetic risk factors only used single SNPs at a time [26,30] or gene expression [27]. Furthermore, a recent study [25] that used machine learning models and incorporated genetic liability only examined one genetic liability at a time. Niu and colleagues [25] utilized three machine learning models, including random forest and neural network, which incorporated a genetic liability component to predict hypertension in rural China. The models included an Asian ancestry hypertension PRS derived from 13 SNPs. The authors observed that integrating hypertension PRS into models improved hypertension incidence prediction and risk classification (AUC random forest = 0.84; AUC neural network = 0.80). Vaura and colleagues [29] incorporated PRS for SBP and DBP in Cox models to assess predictive values of these genetic markers in the risk of incident hypertension prediction using Cox proportional hazards models. The authors observed that including PRS for blood pressure in the clinical prediction model for hypertension increased the C-statistic by 0.5% for the SBP PRS and 0.6% for the DBP PRS. They also observed that incorporating both PRSs in the clinical prediction model resulted in a 0.7% increase in the C-statistic. Our machine learning models incorporated multiple genetic liabilities in the model and showed a 1% improvement in the classification of hypertension. Our research is unique in that it included ten genetic liabilities (incorporating a total of 883 SNPs) utilizing machine learning to establish a more integrated strategy in the classification of hypertension. Our study also took a different approach in terms of the type of genetic liabilities used. Instead of incorporating hypertension genetic liability, we included genetic liabilities for risk factors associated with hypertension and CVD (see methods). The combination of multiple genetic liabilities implemented within machine learning models for the classification of hypertension is the novelty of our work.

Compared with the study by Niu and colleagues, our study achieved a lower performance. Also, in terms of the net reclassification improvement in prediction value, our study showed only a marginal improvement, whereas the study by Niu and colleagues showed an improvement of up to 4.7% in prediction value. This implies that incorporating genetic liabilities relating to the risk factors of hypertension may not be as promising as incorporating the genetic liability of hypertension itself. However, it should be noted that our study investigated a large-scale European ancestry population and the study by Niu, and colleagues investigated a population of Asian ancestry in rural China (The Henan

Rural Cohort Study). These two populations have significant differences in their genetic make-up. Another reason for the observed differences could be environmental exposures and lifestyle variables, which can play a role in modifying the expression or impact of these genetic variants on phenotypes across populations [28,87]. Furthermore, Niu and colleagues did not examine the data-balancing strategy. We employed the data-balancing technique to ensure that our training set was balanced. We also used a feature selection strategy to identify the best features for our machine learning. In addition, we performed model calibration to select the most robust classification model for hypertension.

Our feature selection approach was successful in creating machine learning models that slightly improved the classification of hypertension. However, this came at the price of clinically relevant features (e.g., diabetes mellitus and drinking status) being excluded. We used a specific definition for diabetes (diabetes diagnosed by a doctor, or use of diabetic medication, or Hb1Ac $\geq$ 48 mmol/mol, or glucose level $\geq$ 7.0 mmol/dL) [45]. However, the literature shows that diabetes mellitus and hypertension may co-exist, and it is not exactly clear which of the two precedes the other [88,89]. The observation that our machine learning feature selection approach did not prioritize diabetes as an important feature in classifying hypertension may align with the existing inquiries in the literature regarding the extent to which diabetes influences the development of hypertension or conversely [90].

Our models included lifestyle-related factors, such as BMI, smoking, sedentary lifestyle, age, TC, LDL cholesterol, and HDL cholesterol, as well as genetic liabilities for TC and LDL, which were identified as important features in our best classification model for hypertension. Evidence from the existing literature shows that obesity is accompanied by an increased risk of hypertension due to modifying other risk factors (e.g., elevated levels of LDL cholesterol, reduced levels of HDL cholesterol, and elevated blood pressure) in obese individuals [91]. In addition, high cholesterol levels have been linked to an increased risk of developing hypertension [83]. Furthermore, it has also been shown in previous studies that genetic factors in combination with environmental factors may increase the risk of hypertension and other CVDs [44,92,93]. For example, we have recently shown that physical inactivity in combination with high genetic susceptibility to obesity could increase the risk of hypertension [94]. Our current study did not test for interactions between genetic and lifestyle factors; however, both factors were identified as being important in developing the risk of hypertension.

In our study, we employed random forest and neural network methods, which can effectively capture the hidden interactions between genetic and non-genetic factors in hypertension by leveraging their ability to model nonlinear relationships, handle high-dimensional data, and automatically learn relevant features from the data [62]. These techniques provide excellent tools for examining the complicated aetiology of hypertension and identifying important factors that contribute to its development and progression.

4.1. Strength

A strength of our study is in the novelty of the approaches used including (1) the use of machine learning to build a prediction model of hypertension in a European setting, (2) testing various methods of feature selection to identify the best performing set of predictive features and to ensure that the features included in the final model were robust and that the model was well calibrated, and (3) the addition of multiple genetic liabilities in one single prediction model to identify the best performing classification model. In our integrated genetic approach, we included multiple genetic liabilities comprising a large number of SNPs within ten genetic liabilities and allowed machine learning to identify the best pattern of feature combination in terms of model performance and accuracy. This gave us a comprehensive picture of the effectiveness of various genetic liabilities in comparison with each other and hypertension risk factors. Another strength is in the use of the large sample size of the UKB that allowed us to develop a large training set comprising 171,304 participants. This is beneficial in detecting the true effect of risk factors on outcomes, reducing bias, and making risk predictions in the testing set more reliable [95,96]. Our

study contributes to the ongoing research on the potential role of genetic liabilities in risk prediction of complex diseases [97–99].

4.2. Limitations

A limitation of our research is that the UKB data are imbalanced in terms of the ratio of cases and controls, and, as a result, our sample included 10,528 more controls than cases. The training set included 7370 more controls than cases. Models trained on imbalanced datasets may become biased towards the dominant class, predicting the minority class incorrectly [58]. To address the imbalance in the dataset and minimize error, we utilized an over-sampling approach to balance the sample [100]. The "*ROSE*" package that we used for balancing our data uses an over-sampling approach that may introduce noise into the synthetic sample in the dataset, resulting in some level of bias remaining in the models [101]. In addition, in this study, we employed an integrative approach incorporating multiple features into a machine learning model. To minimize overfitting due to including potentially irrelevant features, we used 5-fold cross-validation techniques using "*caret*" package which allowed us to evaluate the model's performance across multiple training set subsets. However, despite the use of cross-validation techniques, there could still be some residual overfitting in the data due to the model's potential complexity [102]. To mitigate and address the issue of potentially irrelevant features causing overfitting and to improve robustness and generalization as well as performance of our machine learning models, we adopted random forest and neural network, as feature selection techniques, to concentrate on the most significant features to build our machine learning models in stage two.

5. Conclusions

Our research highlighted that out of the ten genetic liabilities examined in our study, genetic liability for two lipid traits (TC and LDL) was found to improve the classification of hypertension within a European population. Incorporating these two genetic liabilities in the random forest model slightly improved hypertension risk discrimination and risk reclassification for participants beyond conventional risk factors.

To improve the generalizability and robustness of classifying hypertension, we propose that future studies incorporate multiple genetic liabilities in machine learning-based models.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jcm13102955/s1, Supplementary Data S1–Supplementary Data S10: List of genetic variants' summary statistics used to construct the genetic risk scores, Supplementary Table S1–Supplementary Table S4, Supplementary Figure S1–Supplementary Figure S7.

Author Contributions: Conceptualization, R.P.; data curation and formal analysis, G.M.; funding acquisition, R.P.; investigation, G.M.; methodology, G.M. and R.P.; project administration, R.P.; resources, R.P.; supervision, R.P; writing—original draft, G.M.; writing—review and editing, G.M. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: Gideon Maccarthy was supported by Brunel University London BRIEF AWARDS 2020/21.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Brunel University London, College of Health, Medicine, and Life Sciences (27684-LR-Jan/2021-29901-1) approve 5 February 2021.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are contained within the article and supplementary materials.

Acknowledgments: This study has been performed using the UKB application 60549.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Available online: https://www.who.int/news-room/fact-sheets/detail/hypertension (accessed on 21 November 2023).
2. Mills, K.T.; Stefanescu, A.; He, J. The global epidemiology of hypertension. *Nat. Rev. Nephrol.* **2020**, *16*, 223–237. [CrossRef] [PubMed]
3. Roth, G.A.; Johnson, C.; Abajobir, A.; Abd-Allah, F.; Abera, S.F.; Abyu, G.; Ahmed, M.; Aksut, B.; Alam, T.; Alam, K.; et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **2017**, *70*, 1–25. [CrossRef] [PubMed]
4. Abdulkader, R.S.; Abera, S.F.; Acharya, D.; Aichour, I.; Aichour, M.T.E.; Akseer, N.; Al-Mekhlafi, H.M.; Aljunid, S.M.; Altirkawi, K.; Ayer, R.; et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **2018**, *392*, 1923–1994.
5. Available online: https://cks.nice.org.uk/topics/hypertension/background-information/prevalence/ (accessed on 22 November 2023).
6. Available online: https://www.gov.uk/government/publications/health-matters-combating-high-blood-pressure/health-matters-combating-high-blood-pressure (accessed on 22 November 2023).
7. Whelton, P.K.; Carey, R.M.; Aronow, W.S.; Casey, J.; Donald, E.; Collins, K.J.; Dennison Himmelfarb, C.; DePalma, S.M.; Gidding, S.; Jamerson, K.A.; et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **2018**, *71*, e127–e248. [PubMed]
8. Schneider, R.; Salerno, J.; Brook, R. 2020 International Society of Hypertension global hypertension practice guidelines—Lifestyle modification. *J. Hypertens.* **2020**, *38*, 2340–2341. [CrossRef] [PubMed]
9. Williams, B.; Mancia, G.; Spiering, W.; Agabiti Rosei, E.; Azizi, M.; Burnier, M.; Clement, D.L.; Coca, A.; de Simone, G.; Dominiczak, A.; et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. *J. Hypertens.* **2018**, *36*, 1953–2041. [CrossRef] [PubMed]
10. Nicoll, R.; Henein, M.Y. Hypertension and lifestyle modification: How useful are the guidelines? *Br. J. Gen. Pract.* **2010**, *60*, 879–880. [CrossRef] [PubMed]
11. Natarajan, P. Polygenic Risk Scoring for Coronary Heart Disease: The First Risk Factor. *J. Am. Coll. Cardiol.* **2018**, *72*, 1894–1897. [CrossRef] [PubMed]
12. Ehret, G.B. Genome-Wide Association Studies: Contribution of Genomics to Understanding Blood Pressure and Essential Hypertension. *Curr. Hypertens. Rep.* **2010**, *12*, 17–25. [CrossRef]
13. Hwang, S.; Vasan, R.S.; O'Donnell, C.J.; Levy, D.; Mattace-Raso, F.U.S.; Morrison, A.C.; Scharpf, R.B.; Psaty, B.M.; Rice, K.; Harris, T.B.; et al. Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* **2009**, *41*, 677–687.
14. Munroe, P.B.; Smith, A.V.; Verwoert, G.C.; Amin, N.; Teumer, A.; Zhao, J.H.; Parsa, A.; Dehghan, A.; Peden, J.F.; Rudan, I.; et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **2011**, *478*, 103–109.
15. Ferreira, T.; Chasman, D.I.; Johnson, T.; Luan, J.; Donnelly, L.A.; Kanoni, S.; Strawbridge, R.J.; Meirelles, O.; Bouatia-Naji, N.; Salfati, E.L.; et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat. Genet.* **2016**, *48*, 1171–1184.
16. Hoffmann, T.J.; Ehret, G.B.; Nandakumar, P.; Ranatunga, D.; Schaefer, C.; Kwok, P.; Iribarren, C.; Chakravarti, A.; Risch, N. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* **2017**, *49*, 54–64. [CrossRef] [PubMed]
17. Warren, H.R.; Evangelou, E.; Cabrera, C.P.; Gao, H.; Ren, M.; Mifsud, B.; Ntalla, I.; Surendran, P.; Liu, C.; Cook, J.P.; et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.* **2017**, *49*, 403–415. [CrossRef] [PubMed]
18. Wang, A.; An, N.; Chen, G.; Li, L.; Alterovitz, G. Predicting hypertension without measurement: A non-invasive, questionnaire-based approach. *Expert Syst. Appl.* **2015**, *42*, 7601–7609. [CrossRef]
19. Kanegae, H.; Oikawa, T.; Suzuki, K.; Okawara, Y.; Kario, K. Developing and validating a new precise risk-prediction model for new-onset hypertension: The Jichi Genki hypertension prediction model (JG model). *J. Clin. Hypertens.* **2018**, *20*, 880–890. [CrossRef] [PubMed]
20. Kanegae, H.; Suzuki, K.; Fukatani, K.; Ito, T.; Harada, N.; Kario, K. Highly precise risk prediction model for new-onset hypertension using artificial intelligence techniques. *J. Clin. Hypertens.* **2020**, *22*, 445–450. [CrossRef] [PubMed]
21. AlKaabi, L.A.; Ahmed, L.S.; Al Attiyah, M.F.; Abdel-Rahman, M.E. Predicting hypertension using machine learning: Findings from Qatar Biobank Study. *PLoS ONE* **2020**, *15*, e0240370. [CrossRef] [PubMed]
22. Zhao, H.; Zhang, X.; Xu, Y.; Gao, L.; Ma, Z.; Sun, Y.; Wang, W. Predicting the Risk of Hypertension Based on Several Easy-to-Collect Risk Factors: A Machine Learning Method. *Front. Public Health* **2021**, *9*, 619429. [CrossRef]
23. Pengo, M.; Montagna, S.; Ferretti, S.; Bilo, G.; Borghi, C.; Ferri, C.; Grassi, G.; Muiesan, M.L.; Parati, G. Machine learning in hypertension detection: A study on world hypertension day data. *J. Hypertens.* **2023**, *41* (Suppl. S3), e94. [CrossRef]

24. Fava, C.; Sjögren, M.; Olsson, S.; Lövkvist, H.; Jood, K.; Engström, G.; Hedblad, B.; Norrving, B.; Jern, C.; Lindgren, A.; et al. A genetic risk score for hypertension associates with the risk of ischemic stroke in a Swedish case–control study. *Eur. J. Hum. Genet.* **2015**, *23*, 969–974. [CrossRef] [PubMed]
25. Niu, M.; Wang, Y.; Zhang, L.; Tu, R.; Liu, X.; Hou, J.; Huo, W.; Mao, Z.; Wang, C.; Bie, R. Identifying the predictive effectiveness of a genetic risk score for incident hypertension using machine learning methods among populations in rural China. *Hypertens. Res.* **2021**, *44*, 1483–1491. [CrossRef]
26. Huang, H.; Xu, T.; Yang, J. Comparing logistic regression, support vector machines, and permanental classification methods in predicting hypertension. *BMC Proc.* **2014**, *8* (Suppl. S1), S96. [CrossRef]
27. Held, E.; Cape, J.; Tintle, N. Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. *BMC Proc.* **2016**, *10* (Suppl. S7), 141–145. [CrossRef] [PubMed]
28. Lu, X.; Huang, J.; Wang, L.; Chen, S.; Yang, X.; Li, J.; Cao, J.; Chen, J.; Li, Y.; Zhao, L.; et al. Genetic Predisposition to Higher Blood Pressure Increases Risk of Incident Hypertension and Cardiovascular Diseases in Chinese. *Hypertension* **2015**, *66*, 786–792. [CrossRef] [PubMed]
29. Vaura, F.; Kauko, A.; Suvila, K.; Havulinna, A.S.; Mars, N.; Salomaa, V.; Gen, F.; Cheng, S.; Niiranen, T. Polygenic Risk Scores Predict Hypertension Onset and Cardiovascular Risk. *Hypertension* **2021**, *77*, 1119–1127. [CrossRef]
30. Li, C.; Sun, D.; Liu, J.; Li, M.; Zhang, B.; Liu, Y.; Wang, Z.; Wen, S.; Zhou, J. A Prediction Model of Essential Hypertension Based on Genetic and Environmental Risk Factors in Northern Han Chinese. *Int. J. Med. Sci.* **2019**, *16*, 793–799. [CrossRef]
31. Albiñana, C.; Zhu, Z.; Schork, A.; Ingason, A.; Aschard, H.; Brikell, I.; Bulik, C.; Petersen, L.; Agerbo, E.; Grove, J.; et al. Multi-PGS enhances polygenic prediction: Weighting 937 polygenic scores. *Nat. Commun.* **2023**, *14*, 4702. [CrossRef]
32. Abraham, G.; Malik, R.; Yonova-Doing, E.; Salim, A.; Wang, T.; Danesh, J.; Butterworth, A.S.; Howson, J.M.M.; Inouye, M.; Dichgans, M. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* **2019**, *10*, 5819. [CrossRef]
33. Krapohl, E.; Patel, H.; Newhouse, S.; Curtis, C.J.; von Stumm, S.; Dale, P.S.; Zabaneh, D.; Breen, G.; O'Reilly, P.F.; Plomin, R. Multi-polygenic score approach to trait prediction. *Mol. Psychiatry* **2018**, *23*, 1368–1374. [CrossRef]
34. Sun, D.; Zhou, T.; Heianza, Y.; Li, X.; Fan, M.; Fonseca, V.; Qi, L. Type 2 Diabetes and Hypertension: A Study on Bidirectional Causality. *Circ. Res.* **2019**, *124*, 930–937. [CrossRef]
35. Giontella, A.; Lotta, L.A.; Overton, J.D.; Baras, A.; Minuz, P.; Melander, O.; Gill, D.; Fava, C. Causal Effect of Adiposity Measures on Blood Pressure Traits in 2 Urban Swedish Cohorts: A Mendelian Randomization Study. *J. Am. Heart Assoc.* **2021**, *10*, e020405. [CrossRef] [PubMed]
36. Miao, K.; Wang, Y.; Cao, W.; Lv, J.; Yu, C.; Huang, T.; Sun, D.; Liao, C.; Pang, Y.; Hu, R.; et al. Genetic and Environmental Influences on Blood Pressure and Serum Lipids Across Age-Groups. *Twin Res. Hum. Genet.* **2023**, *26*, 223–230. [CrossRef]
37. Cadby, G.; Melton, P.E.; McCarthy, N.S.; Giles, C.; Mellett, N.A.; Huynh, K.; Hung, J.; Beilby, J.; Dubé, M.; Watts, G.F.; et al. Heritability of 596 lipid species and genetic correlation with cardiovascular traits in the Busselton Family Heart Study[S]. *J. Lipid Res.* **2020**, *61*, 537–545. [CrossRef] [PubMed]
38. Larsson, S.C.; Mason, A.M.; Bäck, M.; Klarin, D.; Damrauer, S.M.; Million Veteran Program; Michaëlsson, K.; Burgess, S. Genetic predisposition to smoking in relation to 14 cardiovascular diseases. *Eur. Heart J.* **2020**, *41*, 3304–3310. [CrossRef]
39. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **2015**, *12*, e1001779. [CrossRef] [PubMed]
40. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'connell, J.; et al. Genome-wide genetic data on ~500,000 UK biobank participants. *bioRxiv* **2017**. [CrossRef]
41. Welsh, S.; Peakman, T.; Sheard, S.; Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genom.* **2017**, *18*, 26. [CrossRef]
42. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; et al. The Uk biobank resource with deep phenotyping and genomic data. *Nature* **2018**, *562*, 203–209. [CrossRef]
43. Flack, J.M.; Adekola, B. Blood pressure and the new ACC/AHA hypertension guidelines. *Trends Cardiovasc. Med.* **2020**, *30*, 160–164. [CrossRef]
44. Pazoki, R.; Dehghan, A.; Evangelou, E.; Warren, H.; Gao, H.; Caulfield, M.; Elliott, P.; Tzoulaki, I. Genetic Predisposition to High Blood Pressure and Lifestyle Factors: Associations with Midlife Blood Pressure Levels and Cardiovascular Events. *Circulation* **2018**, *137*, 653–661. [CrossRef] [PubMed]
45. Sacks, D.B.; Arnold, M.; Bakris, G.L.; Bruns, D.E.; Horvath, A.R.; Kirkman, M.S.; Lernmark, A.; Metzger, B.E.; Nathan, D.M. Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. *Clin. Chem.* **2011**, *57*, e1–e47. [CrossRef] [PubMed]
46. Mahajan, A.; Taliun, D.; Thurner, M.; Robertson, N.R.; Torres, J.M.; Payne, A.J.; Steinthorsdottir, V.; Scott, R.A.; Grarup, N.; Wuttke, M.; et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **2018**, *50*, 1505–1513. [CrossRef] [PubMed]

47. Winkler, T.W.; Justice, A.E.; Rueeger, S.; Teumer, A.; Ehret, G.B.; Heard-Costa, N.L.; Jansen, R.; Craen, A.J.M.; Boucher, G.; Cheng, Y.; et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLoS Genet.* **2016**, *12*, e1006166. [CrossRef] [PubMed]
48. Shungin, D.; Winkler, T.W.; Croteau-Chonka, D.C.; Ferreira, T.; Locke, A.E.; Mägi, R.; Strawbridge, R.J.; Pers, T.H.; Fischer, K.; Justice, A.E.; et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **2015**, *518*, 187–196. [CrossRef]
49. Liu, M.; Jiang, Y.; Wedow, R.; Brazel, D.M.; Zhan, X.; Agee, M.; Bryc, K.; Fontanillas, P.; Furlotte, N.A.; Hinds, D.A.; et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **2019**, *51*, 237–244. [CrossRef] [PubMed]
50. Surakka, I.; Horikoshi, M.; Mägi, R.; Sarin, A.; Mahajan, A.; Lagou, V.; Marullo, L.; Ferreira, T.; Miraglio, B.; Timonen, S.; et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **2015**, *47*, 589–597. [CrossRef]
51. Marees, A.T.; de Kluiver, H.; Stringer, S.; Vorspan, F.; Curis, E.; Marie-Claire, C.; Derks, E.M. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **2018**, *27*, e1608. [CrossRef]
52. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **2015**, *4*, 7. [CrossRef]
53. Ozsahin, D.U.; Mustapha, M.T.; Mubarak, A.S.; Ameen, Z.S.; Uzun, B. Impact of feature scaling on machine learning models for the diagnosis of diabetes. In Proceedings of the 2022 International Conference on Artificial Intelligence in Everything (AIE), Lefkosa, Cyprus, 2–4 August 2022; The Institute of Electrical and Electronics Engineers, Inc. (IEEE): Piscataway, NJ, USA, 2022.
54. Nguyen, Q.H.; Ly, H.; Ho, L.S.; Al-Ansari, N.; Le, H.V.; Tran, V.Q.; Prakash, I.; Pham, B.T. Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Math. Probl. Eng.* **2021**, *2021*, 4832864. [CrossRef]
55. Pencina, M.J.; D'Agostino, R.B. Evaluating Discrimination of Risk Prediction Models: The C Statistic. *JAMA* **2015**, *314*, 1063–1064. [CrossRef] [PubMed]
56. Xin, J.; Chu, H.; Ben, S.; Ge, Y.; Shao, W.; Zhao, Y.; Wei, Y.; Ma, G.; Li, S.; Gu, D.; et al. Evaluating the effect of multiple genetic risk score models on colorectal cancer risk prediction. *Gene* **2018**, *673*, 174–180. [CrossRef] [PubMed]
57. Nartowt, B.J.; Hart, G.R.; Roffman, D.A.; Llor, X.; Ali, I.; Muhammad, W.; Liang, Y.; Deng, J. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. *PLoS ONE* **2019**, *14*, e0221421. [CrossRef] [PubMed]
58. Kavalci, E.; Hartshorn, A. Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Sci. Rep.* **2023**, *13*, 121. [CrossRef] [PubMed]
59. Wei, Q.; Dunbrack, J.; Roland, L. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE* **2013**, *8*, e67863. [CrossRef]
60. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
61. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *R J.* **2014**, *6*, 79. [CrossRef]
62. Kufel, J.; Bargieł-Łączek, K.; Kocot, S.; Koźlik, M.; Bartnikowska, W.; Janik, M.; Czogalik, Ł.; Dudek, P.; Magiera, M.; Lis, A.; et al. What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics* **2023**, *13*, 2582. [CrossRef]
63. Rajula, H.S.R.; Verlato, G.; Manchia, M.; Antonucci, N.; Fanos, V. Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina* **2020**, *56*, 455. [CrossRef]
64. Montagna, S.; Pengo, M.; Ferretti, S.; Borghi, C.; Ferri, C.; Grassi, G.; Muiesan, M.; Parati, G. Machine Learning in Hypertension Detection: A Study on World Hypertension Day Data. *J. Med. Syst.* **2023**, *47*, 1. [CrossRef]
65. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
66. Wright, M.N.; Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
67. Purkait, N. *Hands-On Neural Networks with Keras: Birmingham*; Packt Publishing: Birmingham, UK, 2019.
68. Islam, M.M.; Alam, M.J.; Maniruzzaman, M.; Ahmed, N.A.M.F.; Ali, M.S.; Rahman, M.J.; Roy, D.C. Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. *PLoS ONE* **2023**, *18*, e0289613. [CrossRef] [PubMed]
69. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002.
70. Ying, X. An Overview of Overfitting and its Solutions. *J. Phys. Conf. Ser.* **2019**, *1168*, 22022. [CrossRef]
71. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
72. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2011.
73. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **2013**, *4*, 627–635.
74. Lindhiem, O.; Petersen, I.T.; Mentch, L.K.; Youngstrom, E.A. The Importance of Calibration in Clinical Psychology. *Assessment* **2020**, *27*, 840–854. [CrossRef] [PubMed]
75. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 621–633. [CrossRef] [PubMed]
76. Steyerberg, E.; Vickers, A.; Cook, N.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.; Kattan, M. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **2010**, *21*, 128–138. [CrossRef]
77. Steyerberg, E.W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*, 2nd ed.; Springer International Publishing: Cham, Switzerland, 2019.

78. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **2010**, *63*, 938–939. [CrossRef]
79. McKearnan, S.B.; Wolfson, J.; Vock, D.M.; Vazquez-Benitez, G.; O'Connor, P.J. Performance of the Net Reclassification Improvement for Nonnested Models and a Novel Percentile-Based Alternative. *Am. J. Epidemiol.* **2018**, *187*, 1327–1335. [CrossRef] [PubMed]
80. Kerr, K.F.; McClelland, R.L.; Brown, E.R.; Lumley, T. Evaluating the Incremental Value of New Biomarkers with Integrated Discrimination Improvement. *Am. J. Epidemiol.* **2011**, *174*, 364–374. [CrossRef] [PubMed]
81. Martens, F.K.; Tonk, E.C.M.; Janssens, A.C.J.W. Evaluation of polygenic risk models using multiple performance measures: A critical assessment of discordant results. *Genet. Med.* **2019**, *21*, 391–397. [CrossRef] [PubMed]
82. Borghi, C.; Veronesi, M.; Bacchelli, S.; Esposti, D.; Cosentino, E.; Ambrosioni, E. Serum cholesterol levels, blood pressure response to stress and incidence of stable hypertension in young subjects with high normal blood pressure. *J. Hypertens.* **2004**, *22*, 265–272. [CrossRef] [PubMed]
83. Wildman, R.P.; Sutton-Tyrrell, K.; Newman, A.B.; Bostom, A.; Brockwell, S.; Kuller, L.H. Lipoprotein Levels Are Associated with Incident Hypertension in Older Adults. *J. Am. Geriatr. Soc.* **2004**, *52*, 916–921. [CrossRef] [PubMed]
84. Ebrahimi, H.; Emamian, M.H.; Hashemi, H.; Fotouhi, A. Dyslipidemia and its risk factors among urban middle-aged Iranians: A population-based study. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2016**, *10*, 149–156. [CrossRef] [PubMed]
85. Xi, Y.; Niu, L.; Cao, N.; Bao, H.; Xu, X.; Zhu, H.; Yan, T.; Zhang, N.; Qiao, L.; Han, K.; et al. Prevalence of dyslipidemia and associated risk factors among adults aged ≥35 years in northern China: A cross-sectional study. *BMC Public Health* **2020**, *20*, 1068. [CrossRef] [PubMed]
86. Wilkinson, I.B.; Prasad, K.; Hall, I.R.; Thomas, A.; MacCallum, H.; Webb, D.J.; Frenneaux, M.P.; Cockcroft, J.R. Increased central pulse pressure and augmentation index in subjects with hypercholesterolemia. *J. Am. Coll. Cardiol.* **2002**, *39*, 1005–1011. [CrossRef]
87. Li, Y.R.; Keating, B.J. Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Med.* **2014**, *6*, 91. [CrossRef]
88. Balogun, W.O.; Salako, B.L. Co-occurrence of diabetes and hypertension: Pattern and factors associated with order of diagnosis among nigerians. *Ann. Ib. Postgrad. Med.* **2011**, *9*, 89–93.
89. Han, L.; Li, X.; Wang, X.; Zhou, J.; Wang, Q.; Rong, X.; Wang, G.; Shao, X. Effect of Hypertension, Waist-to-Height Ratio, and Their Transitions on the Risk of Type 2 Diabetes Mellitus: Analysis from the China Health and Retirement Longitudinal Study. *J. Diabetes Res.* **2022**, *2022*, 7311950. [CrossRef]
90. Petrie, J.R.; Guzik, T.J.; Touyz, R.M. Diabetes, hypertension, and cardiovascular disease: Clinical insights and vascular mechanisms. *Can. J. Cardiol.* **2018**, *34*, 575–584. [CrossRef]
91. Klop, B.; Elte, J.W.F.; Cabezas, M.C. Dyslipidemia in Obesity: Mechanisms and Potential Targets. *Nutrients* **2013**, *5*, 1218–1240. [CrossRef] [PubMed]
92. Tyrrell, J.; Wood, A.R.; Ames, R.M.; Yaghootkar, H.; Beaumont, R.N.; Jones, S.E.; Tuke, M.A.; Ruth, K.S.; Freathy, R.M.; Davey Smith, G.; et al. Gene–obesogenic environment interactions in the UK Biobank study. *Int. J. Epidemiol.* **2017**, *46*, 559–575. [CrossRef]
93. Khera, A.V.; Emdin, C.A.; Drake, I.; Natarajan, P.; Bick, A.G.; Cook, N.R.; Chasman, D.I.; Baber, U.; Mehran, R.; Rader, D.J.; et al. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N. Engl. J. Med.* **2016**, *375*, 2349–2358. [CrossRef]
94. Hezekiah, C.; Blakemore, A.; Bailey, D.; Pazoki, R. Physical activity reduces the effect of adiposity genetic liability on hypertension risk in the UK Biobank cohort. *medRxiv* **2023**. [CrossRef]
95. Biau, D.J.; Kernéis, S.; Porcher, R. Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research. *Clin. Orthop. Relat. Res.* **2008**, *466*, 2282–2288. [CrossRef] [PubMed]
96. Andrade, C. Sample Size and its Importance in Research. *Indian J. Psychol. Med.* **2020**, *42*, 102–103. [CrossRef] [PubMed]
97. Khera, A.V.; Chaffin, M.; Aragam, K.G.; Haas, M.E.; Roselli, C.; Choi, S.H.; Natarajan, P.; Lander, E.S.; Lubitz, S.A.; Ellinor, P.T.; et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **2018**, *50*, 1219–1224. [CrossRef]
98. Khera, A.V.; Chaffin, M.; Wade, K.H.; Zahid, S.; Brancale, J.; Xia, R.; Distefano, M.; Senol-Cosar, O.; Haas, M.E.; Bick, A.; et al. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **2019**, *177*, 587–596.e9. [CrossRef]
99. Yun, J.; Jung, S.; Shivakumar, M.; Xiao, B.; Khera, A.V.; Won, H.; Kim, D. Polygenic risk for type 2 diabetes, lifestyle, metabolic health, and cardiovascular disease: A prospective UK Biobank study. *Cardiovasc. Diabetol.* **2022**, *21*, 131. [CrossRef] [PubMed]
100. Newaz, A.; Mohosheu, M.S.; Al Noman, M.A. Predicting complications of myocardial infarction within several hours of hospitalization using data mining techniques. *Inform. Med. Unlocked* **2023**, *42*, 101361. [CrossRef]
101. Thölke, P.; Mantilla-Ramos, Y.; Abdelhedi, H.; Maschke, C.; Dehgan, A.; Harel, Y.; Kemtur, A.; Mekki Berrada, L.; Sahraoui, M.; Young, T.; et al. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage* **2023**, *277*, 120253. [CrossRef] [PubMed]
102. Lever, J.; Krzywinski, M.; Altman, N. Model selection and overfitting. *Nat. Methods* **2016**, *13*, 703–704. [CrossRef]

Supplementary Material for Chapter 3

Using machine learning to evaluate the value of genetic liabilities in the classification of hypertension within the UK Biobank.

The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jcm13102955/s1, Supplementary Data S1–Supplementary Data S10: List of genetic variants' summary statistics used to construct the genetic risk scores, Supplementary Table S1–Supplementary Table S4, Supplementary Figure S1–Supplementary Figure S7.

Supplementary Table 1: Optimal metrics identified in the random forest models for the classification of hypertension based on the training data.

Model	Number of Features	technique/ set	Outcome	Mtry	OOB Prediction error (Brier s.)
Stage one					
Random forest	10	Random forest /training set	Hypertension	3	0.22
Random forest	20	Random forest /training set	Hypertension	4	0.22
Stage two					
Random forest†	10	Random forest /training set	Hypertension	3	0.22
Random forest††	10	Random forest /training set	Hypertension	3	0.22

†Model built with top ten important features selected by random forest that include conventional risk factors and genetic liabilities. ††Model built with top ten important features selected by neural network that include conventional risk factors and genetic liabilities. Results are from the *ranger* function within the *ranger* R package. Mtry: the number of variables to randomly sample as candidates at each split, OOB: out of bag.

Supplementary Table 2: Optimal metrics identified in the neural network models for the classification of hypertension based on the training data.

Model	Number of Features	Number of Hidden layers	ROC (SD)	Sensitivity (SD)	Specificity (SD)
Stage one					
Neural network	10	5	0.70 (0.004)	0.62 (0.006)	0.67 (0.004)
Neural network	20	5	0.70 (0.004)	0.62 (0.006)	0.67 (0.006)
Stage two					
Neural network †	10	5	0.70(0.004)	0.62(0.007)	0.67(0.006)
Neural network ††	10	5	0.69(0.004)	0.61(0.007)	0.66(0.007)

†Model built with top ten important features selected by random forest that included conventional risk factors and genetic liabilities. ††Model built with top ten important features selected by neural network that included conventional risk factors and genetic liabilities. Results are from the *nnet* function within the *caret* R package. The neural network model was built with parameter for weight decay (decay =0), maximum number of iterations (maxit=100), The maximum allowable number of weights (MaxNWts=1000). ROC: receiver operating characteristic curve, SD: Standard deviation.

Supplementary Table 3: Overview of the association analysis between study predictors and hypertension based on univariable logistic regression analysis.

Characteristic	N	OR	95% CI	*p-value*
Diabetes Mellitus (yes)	244,718	2.26	2.15, 2.38	<0.001
Sex (Male)	244,718	1.55	1.53, 1.58	<0.001
Age	244,718	1.07	1.07, 1.07	<0.001
BMI	244,718	1.12	1.12, 1.12	<0.001
Smoking Status	244,718			
Current	reference	1.0	—	
Never		1.15	1.12, 1.18	<0.001
Previous		1.32	1.28, 1.36	<0.001
Drinking Status	244,718			
Current	reference	1.0	—	
Never		1.12	1.07, 1.17	<0.001
Previous		0.98	0.94, 1.03	0.4
Total Cholesterol	244,718	1.27	1.26, 1.28	<0.001
HDL	244,718	0.72	0.70, 0.73	<0.001
LDL	244,718	1.39	1.37, 1.40	<0.001
Sedentary Lifestyle	244,718	1.07	1.07, 1.07	<0.001

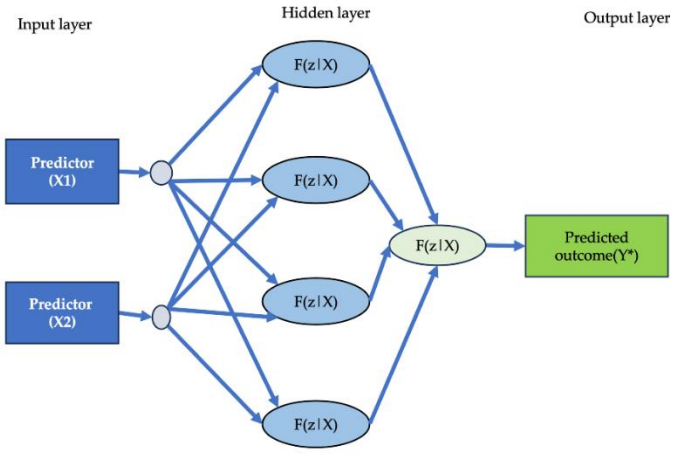
OR: Odds Ratio, CI: Confidence Interval

Supplementary Table 4: Overview of the association analysis between study predictors and hypertension based on multivariable logistic regression analysis.

Characteristic	N	OR	95% CI	*p-value*
Diabetes Mellitus(yes)	244,718	1.56	1.48, 1.65	<0.001
Sex(male)	244,718	1.70	1.67, 1.73	<0.001
Age	244,718	1.07	1.07, 1.08	<0.001
BMI	244,718	1.12	1.12, 1.13	<0.001
Smoking Status	244,718			
Current	reference	1.0	—	
Never		1.11	1.08, 1.14	<0.001
Previous		1.07	1.04, 1.10	<0.001
Drinking Status	244,718			
Current	reference	1.0	—	
Never		0.98	0.94, 1.03	0.5
Previous		0.93	0.88, 0.97	0.003
Total Cholesterol	244,718	1.66	1.59, 1.73	<0.001
HDL	244,718	0.91	0.87, 0.94	<0.001
LDL	244,718	0.63	0.60, 0.66	<0.001
Sedentary Lifestyle	244,718	1.01	1.01, 1.01	<0.001

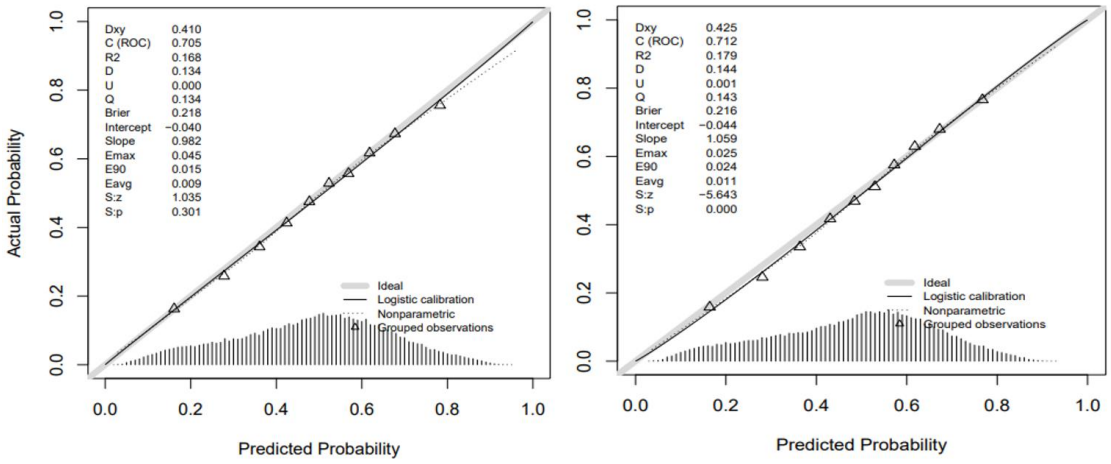
OR: Odds Ratio, CI: Confidence Interval

Supplementary Figure S1:



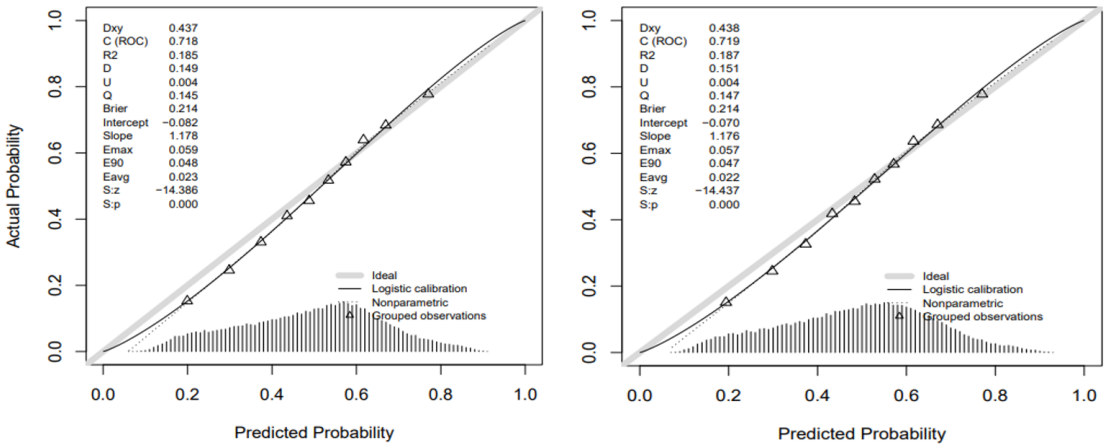
Supplementary Figure S1: A schematic architecture of neural network. F(z|x): Black box function

Supplementary Figure S2



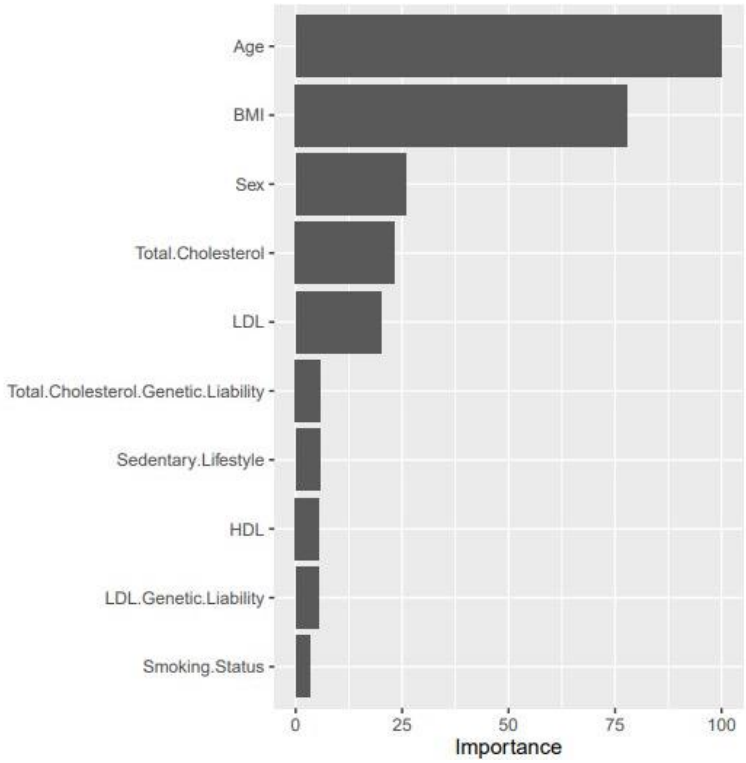
Supplementary Figure S2: Calibration curve of stage one random forest models. The random forest with conventional risk factors (left panel) is well-calibrated. The random forest model that included conventional risk factors and genetic liabilities (right panel) is poorly calibrated due to overfitting. The solid grey line is ideal calibration, the solid black line is logistic calibration, the dotted line is a non-parametric calibration, and the triangular points are the grouped observations. The distribution plot of predicted probability is also displayed.

Supplementary Figure S3



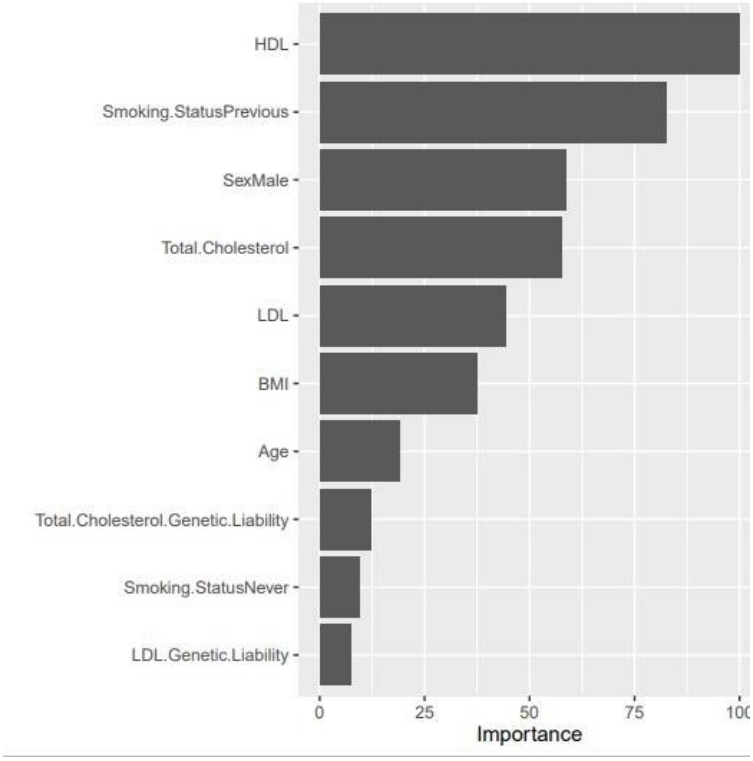
Supplementary Figure S3: Calibration curve of stage one neural network models. The neural network with conventional risk factors (left panel) as well as the neural network model that included conventional risk factors and genetic liabilities (right panel) are both poorly calibrated due to overfitting. The solid grey line is ideal calibration, the solid black line is logistic calibration, the dotted line is a non-parametric calibration, and the triangular points are the grouped observations. The distribution plot of predicted probability is also displayed.

Supplementary Figure S4



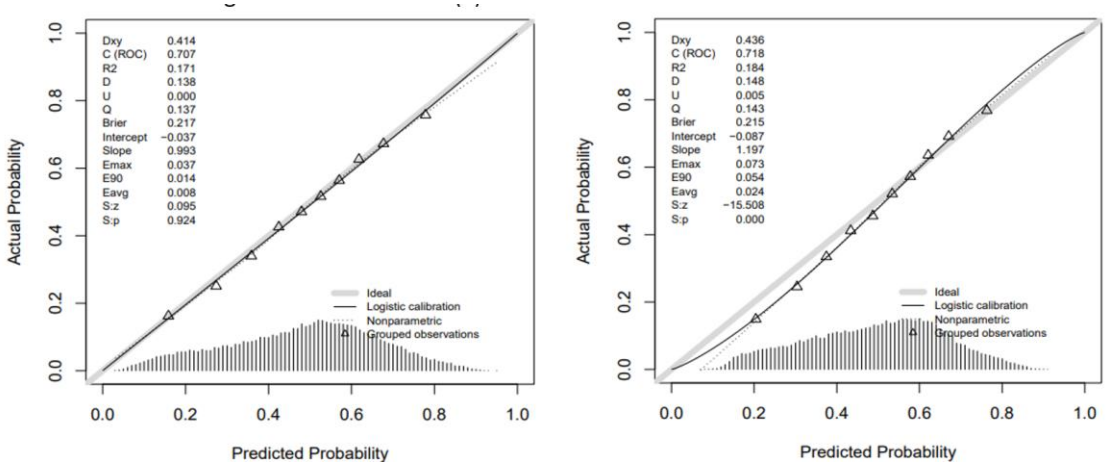
Supplementary Figure S4: Top ten important features selected by the random forest model that include genetic liabilities in addition to conventional risk factors.

Supplementary Figure S5



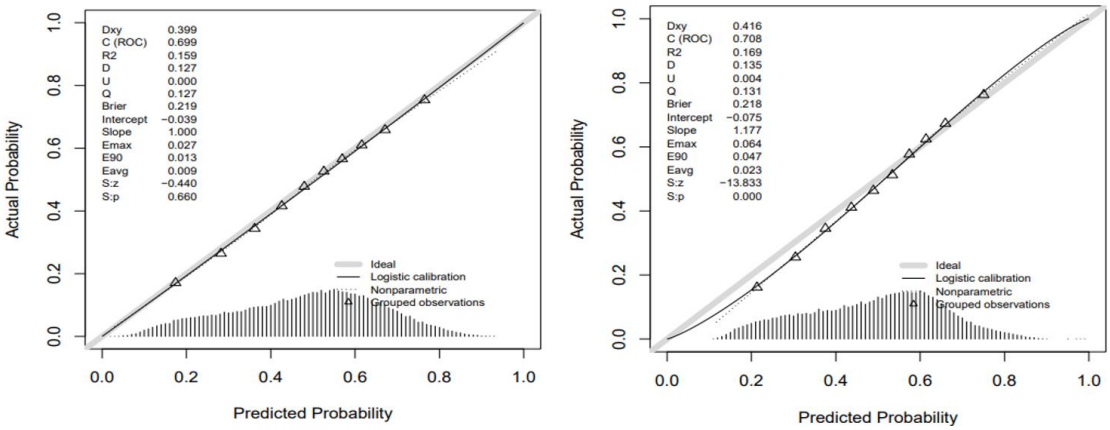
Supplementary Figure S5: Top ten important features selected by the neural network model that include genetic liabilities in addition to conventional risk factors.

Supplementary Figure S6



Supplementary Figure S6: Calibration curve of stage two models created with features selected by random forest model. The model classified with random forest (left panel) is well-calibrated. The model classified with neural network (right panel) is poorly calibrated due to overfitting. The solid grey line is ideal calibration, the solid black line is logistic calibration, the dotted line is a non-parametric calibration, and the triangular points are the grouped observations presence. The distribution plot of predicted probability is also displayed.

Supplementary Figure S7



Supplementary Figure S7: Calibration curve of stage two models created with features selected by neural network model. The model classified with random forest (left panel) is well-calibrated and neural network (right panel) is poorly calibrated due to overfitting. The solid grey line is ideal calibration, the solid black line is logistic calibration, the dotted line is a non-parametric calibration, and the triangular points are the grouped observations presence. The distribution plot of predicted probability is also displayed.

Paper 2

4 Chapter Four. Evaluation of Machine Learning and Traditional Statistical Models to Assess the Value of Stroke Genetic Liability for Prediction of Risk of Stroke Within the UK Biobank.

4.1 Introduction to Paper 2

This paper was published in the Healthcare Journal. It included over 240,000 participants of European ancestry from the UK Biobank. The stroke genetic liability was created using data from MEGASTROKE genome-wide association studies (GWASs). In this study, four predictive models with and without stroke genetic liability were developed, namely a Cox proportional hazard (Coxph) model and a gradient boosting model (GBM), a decision tree (DT), and a random forest (RF), to estimate time-to-event risk for stroke. The study compared the performance of these models' ability to predict the risk of stroke, with a focus on whether incorporating genetic liability to stroke improves prediction accuracy.

healthcare

MDPI

Article

Evaluation of Machine Learning and Traditional Statistical Models to Assess the Value of Stroke Genetic Liability for Prediction of Risk of Stroke Within the UK Biobank

Gideon MacCarthy [1] and Raha Pazoki [1,2,*]

[1] Cardiovascular and Metabolic Research Group, Department of Biosciences, College of Health, Medicine, and Life Sciences, Brunel University of London, Uxbridge UB8 3PH, UK; gideon.maccarthy@brunel.ac.uk
[2] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London W2 1PG, UK
* Correspondence: raha.pazoki@brunel.ac.uk

Abstract: Background and Objective: Stroke is one of the leading causes of mortality and long-term disability in adults over 18 years of age globally, and its increasing incidence has become a global public health concern. Accurate stroke prediction is highly valuable for early intervention and treatment. There is a scarcity of studies evaluating the prediction value of genetic liability in the prediction of the risk of stroke. **Materials and Methods:** Our study involved 243,339 participants of European ancestry from the UK Biobank. We created stroke genetic liability using data from MEGASTROKE genome-wide association studies (GWASs). In our study, we built four predictive models with and without stroke genetic liability in the training set, namely a Cox proportional hazard (Coxph) model, gradient boosting model (GBM), decision tree (DT), and random forest (RF), to estimate time-to-event risk for stroke. We then assessed their performances in the testing set. **Results:** Each unit (standard deviation) increase in genetic liability increases the risk of incident stroke by 7% (HR = 1.07, 95% CI = 1.02, 1.12, *p*-value = 0.0030). The risk of stroke was greater in the higher genetic liability group, demonstrated by a 14% increased risk (HR = 1.14, 95% CI = 1.02, 1.27, *p*-value = 0.02) compared with the low genetic liability group. The Coxph model including genetic liability was the best-performing model for stroke prediction achieving an AUC of 69.54 (95% CI = 67.40, 71.68), NRI of 0.202 (95% CI = 0.12, 0.28; *p*-value = 0.000) and IDI of 1.0×10^{-4} (95% CI = 0.000, 3.0×10^{-4}; *p*-value = 0.13) compared with the Cox model without genetic liability. **Conclusions:** Incorporating genetic liability in prediction models slightly improved prediction models of stroke beyond conventional risk factors.

Keywords: the receiver operation characteristic (ROC); area under the curve (AUC); brier score (BS); integrated calibration index (ICI)

Academic Editor: Joaquim Carreras

Received: 12 February 2025
Revised: 18 April 2025
Accepted: 19 April 2025
Published: 26 April 2025

Citation: MacCarthy, G.; Pazoki, R. Evaluation of Machine Learning and Traditional Statistical Models to Assess the Value of Stroke Genetic Liability for Prediction of Risk of Stroke Within the UK Biobank. *Healthcare* **2025**, *13*, 1003. https://doi.org/10.3390/healthcare13091003

1. Introduction

Stroke is one of the leading causes of mortality and long-term disability in adults over 18 years of age globally [1,2], with a detrimental impact on the economy and the cost of healthcare and social services throughout the world. Stroke survivors have a considerably higher risk of mortality when compared with non-stroke patients, not only attributed to the initial stroke but also to stroke-associated consequences and increased cardiac incidence in years after a stroke [3–6]. Every year, more than 100,000 people in the United Kingdom (UK) suffer from a stroke, and over 1.2 million stroke survivors live in the UK. Stroke incidence

and prevalence in the UK are expected to increase by 60% and 120% annually between 2015 and 2035, respectively [7].

Studies have shown that both genetic and non-genetic factors play a critical role in the complex process of stroke events [8]. Stroke risk increases with age, with an estimated 10-year stroke risk in those aged 55 and over. The risk varies by gender and the increasing co-occurrence of risk factors, such as hypertension, diabetes mellitus, atrial fibrillation, high blood cholesterol and lipids, cigarette smoking, physical inactivity, chronic kidney disease, and family history [9].

Twin and family history studies provided early evidence that genetics had a role in stroke risk [10]. Genome-wide association studies (GWASs) have provided further evidence to confirm the role of genetic factors in the occurrence of stroke. More recently, large-scale GWASs, such as the International Stroke Genetics Consortium (ISGC), have identified genetic loci associated with stroke. The MEGASTROKE project identified over 32 loci contributing to stroke risk, revealing the causal role of specific genes and gene regions in stroke origins [11,12]. As a result, greater insight into the genetic indicators of stroke has allowed an opportunity for a deeper evaluation of an individual's stroke risk, as well as potentially more informed medical and lifestyle decisions that may be preventative measures to reduce the risk of stroke occurrence.

Prediction tools for stroke, such as the Framingham Stroke Risk Profile (FSRP), the American Heart Association (AHA), and the American Stroke Association (ASA), are critical in identifying at-risk individuals early on, allowing for timely treatments and improving outcomes [13,14]. Their developments extend beyond individual treatment, including healthcare policy, budget allocation, and ethical issues for patient data. Advances in artificial intelligence and machine learning are pushing the boundaries of prediction tools, making them more accurate and adaptive to diverse groups of patients [15].

The stroke prediction tools (FRSPs and ASA), as well as the current clinical guidelines for cardiovascular disease prevention, do not evaluate or integrate genetic liability into the risk assessment [16]. Genetic prediction of stroke has the potential to transform stroke prevention and treatment. It has the potential to identify individuals who are at risk of or predisposed to stroke even before clinical symptoms appear. This allows for early treatments, such as lifestyle adjustments or personalized drug programs [17–21]. Genetic polymorphisms in genes associated with stroke or its risk factors have been investigated in stroke risk. Several studies reported a significant association with stroke risk and a genetic liability derived from a set of single nucleotide polymorphisms (SNPs) that were previously identified to have a strong association with stroke or stroke risk factors [22–28]. Genome-wide genetic liabilities, derived from the combined effects of several genetic variants across the genome, regardless of the strength of their association, have been increasingly tested in the last decades for their effect in health and disease, and previous studies have shown that higher scores of genome-wide genetic liabilities enhance the stroke risk prediction [29,30].

Machine learning models are increasingly applied to predict the risk of complex diseases [31–42]. Studies focusing on the prediction of the risk of stroke [32,41,42] have shown that machine learning models outperformed traditional statistical techniques, such as the Cox proportional hazards model. However, there is no consistency on which machine learning model is a better fit. Chen et al. [42] identified artificial neural networks (ANNs), whilst Chun et al. [32] found that gradient-boosted trees (GBTs) were superior to other machine learning models. In addition, Wang et al. [41] identified that the random forest approach outperformed the Cox proportional hazards model.

The predictive value of the genetic factors used in machine learning models is unclear. In a case–control study focusing on patients with atrial fibrillation, Papadopoulou et al. [40] showed that out of multiple machine learning models incorporating a genetic liability, XGBoost

outperformed a widely used existing clinical prediction model (CHA2DS2-VASc). The study by Papadopoulou et al. [40] did not include incident stroke, and they created their genetic liability using a selected list of SNPs associated with ischemic stroke (Supplementary Table S1).

To our knowledge, there is currently no study in the European general population that provides a comprehensive insight into the prediction of the risk of incident stroke in various scenarios, incorporating machine learning and a stroke genome-wide genetic liability. To fill this gap, our research focused on incorporating a genome-wide genetic liability into machine learning for the prediction of the risk of incident stroke using survival data. This would offer a better understanding of the additional benefit of genetic liability in stroke risk prediction, as well as of how machine learning algorithms perform in comparison to traditional survival models in this context.

We have three main objectives, including (1) assessing the association of whole-genome liability and the risk of future stroke occurrence (incident stroke), (2) assessing the predictive value of stroke genetic liability in the prediction of stroke, and (3) comparing the performance of the Cox proportional hazard model and machine learning models before and after incorporating genome-wide stroke genetic liability into the model.

2. Material and Method

2.1. Ethical Approval

The Northwest Multi-Centre Research Ethics Committee approved the UK Biobank (UKB) as a research tissue bank, and all participants involved in the UKB project provided informed consent. The current study is based on UKB data, with the application number 60549. In addition, Brunel University of London's College of Health, Medicine and Life Sciences Research Ethical Committee approved the use of UKB secondary data (reference 27684-LR-Jan/2021-29901-1).

2.2. Study Population

The UK Biobank (UKB) is a prospective observational research study including more than 500,000 adults aged between 40 and 69 years. From 2006 to 2010, participants were recruited from 22 centres across the United Kingdom. The comprehensive description of the UK Biobank study, the acquired data, and a summary of its characteristics are publicly accessible on the UK Biobank website (www.biobank.ac.uk, viewed on 20 June 2021) and in other sources, including Sudlow et al. [43]. During the recruitment stage, detailed information about socioeconomics, demographics, health status, family history of diseases, and lifestyle variables was obtained from the participants through questionnaires and interviews. Several physical measurements were obtained, including height, weight, body mass index (BMI), waist–hip ratio (WHR), systolic blood pressure (SBP), and diastolic blood pressure. The records of UKB study participants were linked to health episode statistics (HES) data and national death and cancer registries.

The current study focuses on a sample of unrelated participants of European ancestry (N = 243,399; Figure 1). In brief, we employed 40 genetic principal components developed centrally by the UKB and used the k-means clustering technique on 502,219 UKB participants to identify persons of European ancestry who had available genetic data (N = 459,042). The study eliminated participants who had withdrawn their informed consent (N = 61), pregnant women, and those who were uncertain about their pregnancy status (N = 278). We excluded participants whose self-reported sex did not match their genetic sex (n = 320). We excluded people who were first and second-degree relatives (N = 33,369) by using a kinship cutoff of 0.0884 for third-degree relatives. We removed individuals (N = 25,340) who had been diagnosed with vascular or cardiac issues by a clinician before or during recruitment. This was carried out to minimize possible confounding, the influence of re-

verse causality, and selection biases. Participants who used cholesterol-lowering medicine (N = 34,243), quit smoking or drinking due to health reasons or doctor's advice (n = 58,752), or had missing data on confounders (N = 61,961) were also removed from the dataset.

We subsequently excluded participants who had prevalent stroke cases (N = 248), and self-reported stroke (N = 130). We then merged the data with genetic liability profile data (N = 425,054) calculated for participants with available genotype data (N = 459,042), leaving a final 243,399 unrelated individuals of European ancestry.

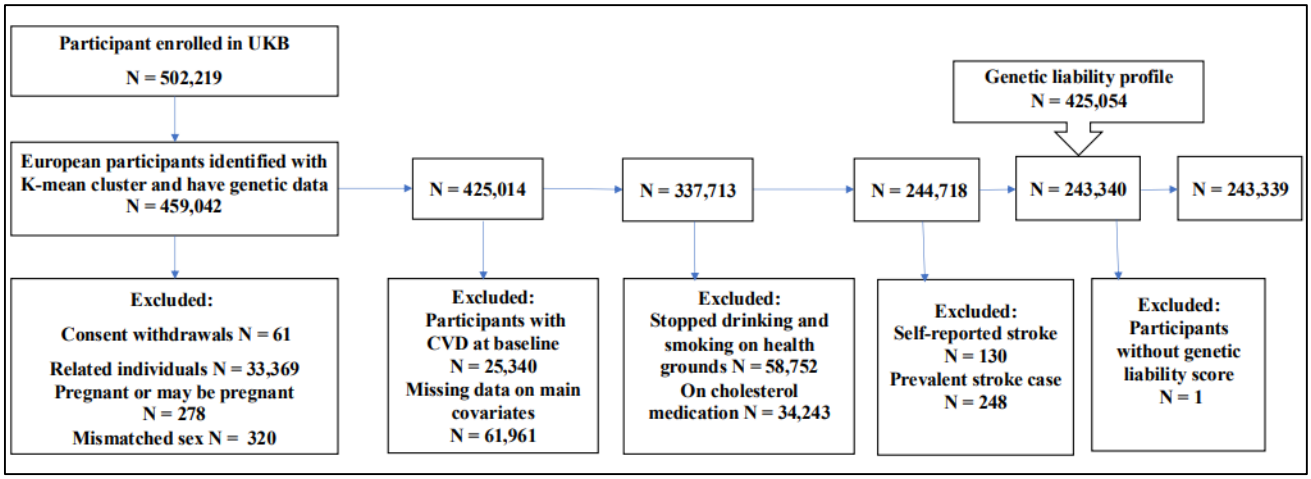


Figure 1. Exclusion criteria of the study: The flowchart for selecting research participants. At the start of this study, the UK Biobank (UKB) had over 500,000 participants. We employed the K-means cluster approach to extract 459,042 European-ancestry subjects. The final dataset had 243,339 people who satisfied the inclusion criteria.

2.3. Genotyping and Imputation

The UKB conducted all DNA extraction, genotyping, and imputation. Details of procedures are discussed elsewhere [44–46]. To summarize, blood samples from participants were taken at UKB assessment centers, and DNA was extracted and genotyped using the UKB Axiom array. UKB used the IMPUTE4 program [47] to perform the genotyping imputation. The three reference panels used for imputation were the Haplotype Reference Consortium, UK10K, and 1000 Genomes Phase 3. The UKB generated genetic principal components and kinship coefficients centrally to identify related individuals and adjust for population stratification [44,46].

2.4. Definition of the Outcome

Our primary outcome in the current study was stroke events, defined according to the International Classification of Diseases 10th revision (ICD-10, I60–I67). In this study, incident stroke was characterized using cerebrovascular disorders ICD-10 code (I600–I609, I610–I619, I630–I639, I64, I650–I659, I660–I669, and I670–I679) for the first stroke event. The current study's follow-up period is computed from the date of health assessment upon enrolment to the end of March 2017. The participants who did not experience the outcome at the end of the follow-up period were censored.

2.5. Demographics and Clinical and Lifestyle Features

In this study, the conventional risk factors, including age, sex, BMI, diabetes mellitus (DM), hypertension, total cholesterol (TC), low-density lipoprotein (LDL), smoking, and drinking, were considered in all the analyses. A doctor's diagnosis of diabetes, the usage of insulin, a blood hemoglobin (HbA1c) level greater than or equal to 48 mmol/mol (6.5%), or a glucose level greater than or equal to 7.0 mmol/dL were all considered indicators

of diabetes mellitus (DM) [48]. Hypertension is defined as (1) having a recorded SBP greater than or equal to 140 mmHg or DBP greater than or equal to 90 mmHg, (2) having a doctor-diagnosed case of hypertension, or (3) having a record of taking blood pressure (BP)-lowering medication at baseline [49,50].

In the UKB, a manual sphygmomanometer or a standard automated device was used to collect two blood pressure readings, separated by a few minutes (https://biobank.ctsu.ox.ac.uk/ukb/ukb/docs/Bloodpressure.pdf (accessed on 22 November 2021)). Using two automatic or two manual blood pressure readings, we calculated the mean SBP and mean DBP. The average of the two values was used for people who had one manual and one automated blood pressure reading. For participants having a single blood pressure record, that one blood pressure reading was used for those participants. For participants using blood pressure-lowering drugs, we increased SBP by 15 mmHg and DBP by 10 mmHg [51]. We excluded individuals with incomplete blood pressure readings from the study. The UKB used a self-reported questionnaire to collect data on participant smoking and alcohol consumption, and categorized respondents as never, previous, and current consumers.

2.6. Computation of Genetic Liabilities

Selection of Genetic Variants

We selected a list of genetic variations, in the form of SNPs (Supplementary Data S1), that were previously identified in the European population as being associated with stroke [11]. The effect sizes for these SNPs (Supplementary Data S1) were derived from GWAS summary statistics data that were published and made publicly available on the GWAS Catalog website (https://www.ebi.ac.uk/gwas/, visited on 12 July 2021). SNPs with a minor allele frequency (MAF) of less than or equal to 0.01 and duplicate, non-biallelic SNPs were not included in the genetic liability calculation for this study. We also conducted an LD pruning technique to exclude SNPs that were in linkage disequilibrium (LD) with one another. When the correlation between SNPs occurs more frequently than expected in a random sample, the SNPs are said to be in LD [52]. LD between two loci is statistically determined by using metrics, such as the correlation coefficient (r^2) value. This value measures how well the alleles at the two loci correlate with one another. LD pruning removes highly correlated SNPs to avoid the statistical bias and computational inefficiency caused by LD. For this LD pruning process, all pairs of SNPs within a given moving window are evaluated to determine their pairwise LD based on r^2 value. If any pair of SNPs within the window has an LD larger than the stated threshold, the first SNP will be pruned [53]. The pruning process was implemented in PLINK version 1.9 [54] with the function and parameters "*--indep-pairwise window size = 250 step size = 50 r^2 = 0.1*". After the LD pruning procedure, 252,903 SNPs were retained for calculating the genetic liability for stroke based on the Purchell method [54] (Supplementary Figure S1).

The calculation of genetic liability for stroke was implemented in PLINK version 1.9 with the function "*-score*". PLINK employs a weighted technique in which the effect size (beta coefficient) of each SNP is used as a weight and is multiplied by the number of risk alleles carried by the participant. The result is then summed up across all SNPs in the calculation of genetic liability.

2.7. Data Preprocessing

We preprocessed the dataset by standardizing all quantitative variables, including age, BMI, TC, LDL, and genetic liability using the "*scale*" function in the R package. Categorical variables included sex (male and female), smoking status (never, previous, current), alcohol consumption status (never, previous, current), DM (no, yes), and hypertension (no, yes).

Genetic liability was additionally categorized as low, medium, and high risk according to its tertiles to ease the analysis per subgroup of genetic liability.

3. Statistical Analysis

For a statistical description of the baseline characteristics of our study population, we used the "*gtsummary*" and "*table1*" packages in the R-program Windows version 4.4.1 for statistical analyses [55]. The categorical variables were summarized using frequencies and percentages, and the numerical variables were expressed as the mean (SD). The chi-square test was used to compare differences in binary outcome (stroke event and non-event) in relation to categorical variables. For continuous variables, the Wilcoxon rank sum test was used. We used the "*cor*" function to calculate the correlation matrix and the Pearson correlation between variables and the "*ggcorrplot*" function from the *ggcorrplot* package to visualize the correlation matrix. We then examined the correlation matrix using the "*findCorrelation*" function from the *caret* package to identify highly correlated features. In this study, we set the Pearson correlation ($r^2 = 0.8$) as the threshold for collinearity [56,57].

The feature selection procedure began with (1) selecting risk factors known to be associated with stroke and (2) were associated with stroke in our data using the univariate Cox regression (*p*-value less than 0.05 for inclusion). (3) We then used the correlation coefficient to assess the correlation among the selected risk factors (r^2 less than 0.8 for inclusion). The finally selected risk factors were used to construct the conventional risk factor model (model 1).

3.1. The Relationship Between Genetic Liability and Stroke

We used univariable and multivariable Cox proportional hazard regression to assess the relationship between stroke genetic liability (continuous and categorical) and the risk of incident stroke over the follow-up period. Hazard ratios (HRs) are commonly used to evaluate outcomes, such as survival time and time to event. HR is a measure used in survival analysis to compare the risk of an event occurring at any given point in time between two groups.

Following the univariable Cox proportional hazard regression analysis (model 1, unadjusted), three multivariable adjustment Cox proportional hazard regression models (models 2, 3, and 4) were developed to examine the potential influence of known cardiovascular risk factors on the relationship between genetic liability and stroke risk. In model 2, we adjusted for age and sex. In model 3, BMI, hypertension, DM, and LDL were adjusted in addition to age and sex, and in model 4, we further adjusted for drinking status and smoking status (the full model). We identified statistical significance when the associations established a two-sided *p*-value less than 0.05. We assessed the proportional hazard (PH) assumptions using statistical testing (the "*cox.zph*" function) and a visual examination of scaled Schoenfeld residuals (the "*ggcoxzph*" function) using the R *survival* package version 3.8-3.

3.2. Prediction Models Development

In this study, two sets of prediction models were created for each technique to predict the incidence of stroke. These were (1) the conventional risk factors model (the model without genetic liability), which combines the conventional risk factors selected from univariable association tests, and (2) the integrated prediction model, which combines the conventional risk factors with genetic liability for stroke (genetic risk). The input variables and output in the current study are displayed in Supplementary Figure S2.

Using the "*createDataPartition*" function from the *caret* package, we randomly partitioned our dataset into a training set (70%; N = 170,381; event = 1382; non-event = 168,999) and a testing set (30%; N = 73,018; event = 591; non-event = 72,427).

To predict the risk of incident stroke, we used the training data to create prediction models using the Cox proportional hazard. Cox proportional hazard regression [58,59] is a popular statistical approach for assessing survival data and determining the association between the time until an event (such as death, failure, or illness recurrence) occurs and one or more predictors. We implemented the Cox proportional hazard models using the "*coxph*" function from the *Survival* package in R software version 3.8-3.

In addition, we developed three machine learning techniques in the training set, including the gradient boosting machine (GBM) models, decision tree (DT), and random forest (RF), to predict the risk of stroke. We then assessed the performance of each model in the testing set (Supplementary Figure S3).

The decision tree is one of the common and simple methods used for classification and regression applications. It works by dividing a dataset into smaller subgroups depending on feature values and then generating a decision tree [60]. The decision tree method in this study was implemented using the "*rpart*" function from the recursive partitioning and regression trees (*rpart*) package, and the minimum number of observations required to split a node at each branch was set to 4. The complexity parameter (cp) to control the size of the decision tree and prevent overfitting was set at 0.001, meaning that a split must improve the model's fit by at least 0.1% to be considered. This parameter is used to save computing time by removing irrelevant splits. The optimal decision tree was obtained with the "*prune*" function. The function removes the trees that do not meet the complexity parameter value. That is, the "*prune*" function removes branches without a lack of fit reduction (measured by the residual sum of squares; RSS) as determined by the complexity parameter value. This process reduces the risk of overfitting the training data.

Random forest is a popular machine learning model for classification and regression. It creates ensembles from decision trees and combines their results to make a final decision [61]. The random forest models were built using the "*ranger*" function from the *ranger* package. The number of trees to be fitted was set to a value of 500. To control the model's complexity and performance, the number of variables randomly selected at each split when growing the trees was set to a value of 3 ("*mtry*"). This is justified, as the optimal *mtry* value considered for classification models is calculated as the square root of the total number of variables (nine variables in the current study). The value of *mtry* can significantly affect the OOB (out-of-bag) error. The OOB error is an unbiased estimate of the prediction error calculated by using samples not included in the bootstrap sample for a given tree. It serves as a cross-validation mechanism that is integrated into the random forest. A smaller *mtry* value increases the randomness and diversity among the trees, which can help reduce overfitting and potentially lower the OOB error. However, if the *mtry* value is too small, the trees might not capture enough information, leading to higher OOB error. The *mtry* and OOB error are critical in optimizing the random forest model. The range of values for *mtry* was examined by the *ranger* package version 0.17.0, and the *mtry* value that minimizes OBB error was selected as the optimal value in the construction of the random forest model. We additionally built gradient boosting machine models using the *gbm* package version 2.2.2 to predict the risk of stroke. The gradient boosting machine models integrate predictions from many weak learners to increase total prediction accuracy [60]. The number of trees to be fitted was set to a value of 500. The highest number of permissible variable interactions was set to 3. The shrinkage parameter to control the learning rate or step-size reduction was set to a value of 0.01. The parameters of the machine learning models were determined using 10-fold cross-validation (CV). In this study, the parameters with the smallest CV root mean square error (RMSE), CV error (xerror), and OOB error were utilized to develop the GBM, DT, and RF prediction models, respectively.

4. Model Performance Assessment

To determine the predictive performance of each prediction model, we used the Platt scaling method [62], also known as the sigmoid method, which is commonly used in machine learning methods for binary data. This method calibrates the output of the prediction models. Platt scaling transforms the output from classification models into a probability distribution. Here, we passed the probability estimates from machine learning models through a trained sigmoid function [62] using univariable logistic regression. In this logistic regression, a variable containing probability estimates for each participant was used as an independent variable. The binary outcome (stroke) served as the dependent variable [63]. The output from this logistic regression provided a new scaled probability estimate that helped calibrate the models. The calibration of a prediction model ensures that the predicted risks are accurate and align with the actual proportions of the event. A prediction model is said to be calibrated if the model's outcome matches the observed proportions of the event [64]. To assess the agreement between the calibrated probabilities (created using Platt scaling) and the observed patient stroke outcomes, we additionally used the "*pmcalibration*" function from *pmcalibration* in R package. This method allows for nonlinear relationships between the predictors and the response variables. Complementary log–log transformed predicted probabilities were applied to the splines to produce calibration measures for a time-to-event outcome.

The calibration metrics used to assess the model calibration in this study were the Brier score (*BS*) and average absolute difference (*Eavg*), also known as the integrated calibration index (*ICI*). The BS is the mean squared difference between the predicted probabilities and the actual outcomes, and it measures both discrimination and calibration [63]. BS ranges from 0 (perfect prediction and calibration) to 1 (worse prediction and calibration). ICI measures the average absolute deviation between the predicted and observed probabilities, providing an overall assessment of calibration quality [64]. It provides a single, summary measure of calibration quality, making it easier to compare different models or assess changes in calibration over time. An ICI of 0 represents perfect calibration and an ICI of 1 represents worse calibration, suggesting that the predicted probability deviates from the observed events. To calculate ICI and BS, we used the "*pmcalibration*" and "*brier*" functions implemented within the *pmcalibration* and *gmish* packages, respectively.

To assess the discrimination performance of the models, we calculated the area under the curve (AUC) using the *pROC* package in the R program. We reported the AUC, ICI, and BS values of various models. Greater values of AUC and smaller values of ICI and BS indicate improved discrimination and calibration of the model. The overview of the model performance assessment is presented in Supplementary Figure S3.

Assessment of the Predictive Value of Genetic Liability

We assessed the predictive value of genetic liability as an additional predictor to the conventional risk factors in each prediction model by estimating the improvement in the AUC, integrated discrimination improvement (IDI), and continuous net reclassification index (NRI). NRI measures the effectiveness of a new model in reclassifying individuals into different risk categories compared to an existing model. At the same time, IDI evaluates the model's ability to differentiate between cases and non-cases after adding a new variable. It compares the average predicted probability for cases and non-cases in the old and new models [65]. The NRI and IDI were calculated to assess model improvement following the inclusion of genetic liability in the models. This was implemented using the "reclassification" function from the *PredictABEL* package version 1.2-4 in the R-program. Higher IDI value indicated better discrimination, and higher NRI value indicated better

risk reclassification by the new model [66–68]. The above performance metrics have been discussed in detail, elsewhere [63] and in our previous work [33].

5. Results

5.1. Study Characteristics

Table 1 presents the baseline characteristics of the study. The study included 243,339 unrelated UK Biobank participants of European ancestry. The average age of participants included in the study was 55.4 (SD = 7.98) years at recruitment. Over half of the sample were women (N = 141,212; 58%). During a median follow-up of 8.22 years, 1973 first-ever stroke episodes, of which 45.3% of patients were women, were recorded among the participants.

In the overall sample, 76,397 participants (31.4%) were current smokers, and 228,349 participants (93.8%) were current alcohol drinkers. All the conventional risk factors included in the analysis showed a statistically significant association with the risk of incident stroke in univariate analysis except total cholesterol (Table 1). The prevalence of DM within the sample was 2.9% (N = 6939) while the prevalence of hypertension was 47.8% among the participants (N = 116,216). The univariable Cox association analysis results indicated that age, sex, BMI, hypertension, DM, LDL, alcohol use, and smoking history were statistically associated with the risk of stroke (Table 1). These variables were used as the features to construct conventional risk factor models. The correlation matrix (Figure 2) between the characteristics in the study demonstrated that total cholesterol and LDL were highly correlated (r^2 = 0.94). LDL was used in the further analysis and feature selection.

Table 1. Baseline characteristics of the study population stratified for stroke event and non-stroke event within the UK Biobank population.

Characteristic	Overall (N = 243,399)	Non-Event (N = 241,426)	Stroke Event (N = 1973)	HR (95% CI)	*p*-Value
DM, yes; n (%)	6939 (2.9%)	6826 (2.8%)	113 (5.7%)	2.08(1.72, 2.51)	<0.001
Hypertension, yes; n (%)	116,216 (47.7%)	114,840 (47.6%)	1376 (69.7%)	2.52(1.29, 2.78)	<0.001
Sex, male; n (%)	102,187 (42.0%)	101,107 (41.9%)	1080 (54.7%)	1.67 (1.53, 1.83)	<0.001
Age (years), mean (SD)	55.4 (7.98)	55.4 (7.98)	60.0 (7.14)	1.93 (1.83, 2.03)	<0.0001
Body mass index (kg/m^2), mean (SD)	26.8 (4.57)	26.8 (4.57)	27.4 (4.83)	1.12 (1.08, 1.17)	<0.001
Total cholesterol (mmol/L), mean (SD)	5.91 (1.06)	5.91 (1.06)	5.94 (1.09)	1.03 (0.98, 1.07)	0.30 *
LDL (mmol/L), mean (SD)	4.68 (2.37)	4.67 (2.36)	5.03 (2.51)	1.03 (1.03, 1.12)	0.002
Smoking					
Current; n (%)	76,397 (31.4%)	75,647 (31.3%)	750 (38.0%)	REF	REF
Previous; n (%)	2900 (1.2%)	2855 (1.2%)	45 (2.3%)	1.58 (1.17, 2.13)	0.003
Never; n (%)	164,102 (67.4%)	162,924 67.5%)	1178 (59.7%)	0.73 (0.67, 0.80)	<0.001
Alcohol					
Current; n (%)	228,349 (93.8%)	226,556(93.8%)	1793 (90.9%)	REF	REF
Previous; n (%)	7082 (2.9%)	6996 (2.9%)	86 (4.4%)	1.55 (1.25, 1.93)	<0.001
Never; n (%)	7968 (3.3%)	7874 (3.3%)	94 (4.8%)	1.50 (1.22, 1.85)	<0.001

The *p*-value is from a univariate analysis of the Cox proportional hazard model, comparing the distribution of the baseline characteristics among stroke and non-stroke event. * Not significant; DM = diabetes mellitus; HR = hazard ratio; CI = confidence interval; REF: reference.

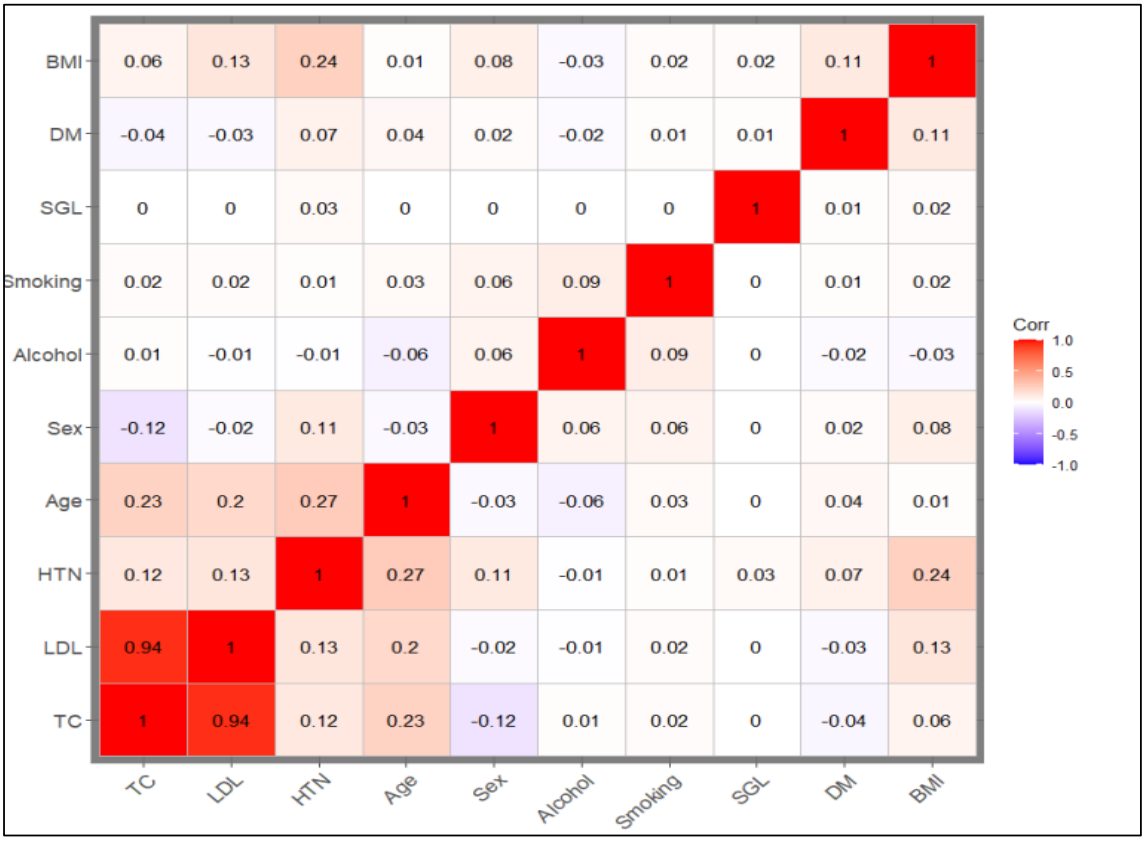


Figure 2. Correlation matrix plot: The plot shows the correlation coefficients between numerical features. TC and LDL are highly correlated ($r^2 > 0.8$). TC was excluded from further analysis (prediction model construction). BMI: body mass index; TC: total cholesterol; LDL: low-density lipoprotein cholesterol; HTN: hypertension; SGL: stroke genetic liability.

5.2. The Association of Genetic Liability with Incident Stroke

The Kaplan–Meier curve showed differences in stroke incidents and cumulative hazard between the high-risk and low-risk genetic liability groups (Figure 3). Each unit (standard deviation) increase in genetic liability increases the risk of incident stroke by 7% (HR = 1.07, 95% CI = 1.02, 1.12, *p*-value = 0.003; Table 2; Figure 4).

The risk of stroke was greater in the higher genetic liability group, as demonstrated by a 14% increased risk (HR = 1.14, 95% CI = 1.02, 1.27, *p*-value = 0.02) compared with the low genetic liability group. The global Schoenfeld *p*-value from the Schoenfeld test (*p*-value = 0.14; Table 3) indicates that the proportional hazard (PH) assumption is reasonable for the model (Supplementary Figure S4).

Table 2. The result of the univariable Cox proportional hazard model for the association of genetic liability (categorical and continuous) with incident stroke within the UK Biobank population.

Genetic liability Level	HR (95% CI)	*p*-Value	HR (95% CI)	*p*-Value	HR (95% CI)	*p*-Value	HR (95% CI)	*p*-Value
	Model 1		**Model 2**		**Model 3**		**Model 4**	
Moderate risk	1.06 (0.95, 1.18)	0.31	1.06 (0.95, 1.18)	0.31	1.05 (0.94, 1.17)	0.04	1.05 (0.94, 1.17)	0.40
High risk	1.15 (1.03, 1.28)	0.01	1.16 (1.04, 1.30)	0.01	1.14 (1.02, 1.27)	0.02	1.14 (1.02, 1.27)	0.02
Genetic liability (continuous)	1.08 (1.03, 1.13)	<0.001	1.08 (1.03, 1.13)	<0.001	1.07 (1.03, 1.12)	0.002	1.07 (1.02, 1.12)	0.003

Model 1: univariable Cox proportional hazard. Model 2: adjusted for age and sex. The low genetic risk group was considered as the reference. Model 3: adjusted for age, sex, BMI, LDL, HTN, and DM. Model 4: adjusted for age, sex, BMI, LDL, HTN, DM, smoking status, and alcohol status. BMI: body mass index; LDL: low-density lipoprotein cholesterol; HTN: hypertension; DM: diabetes mellitus.

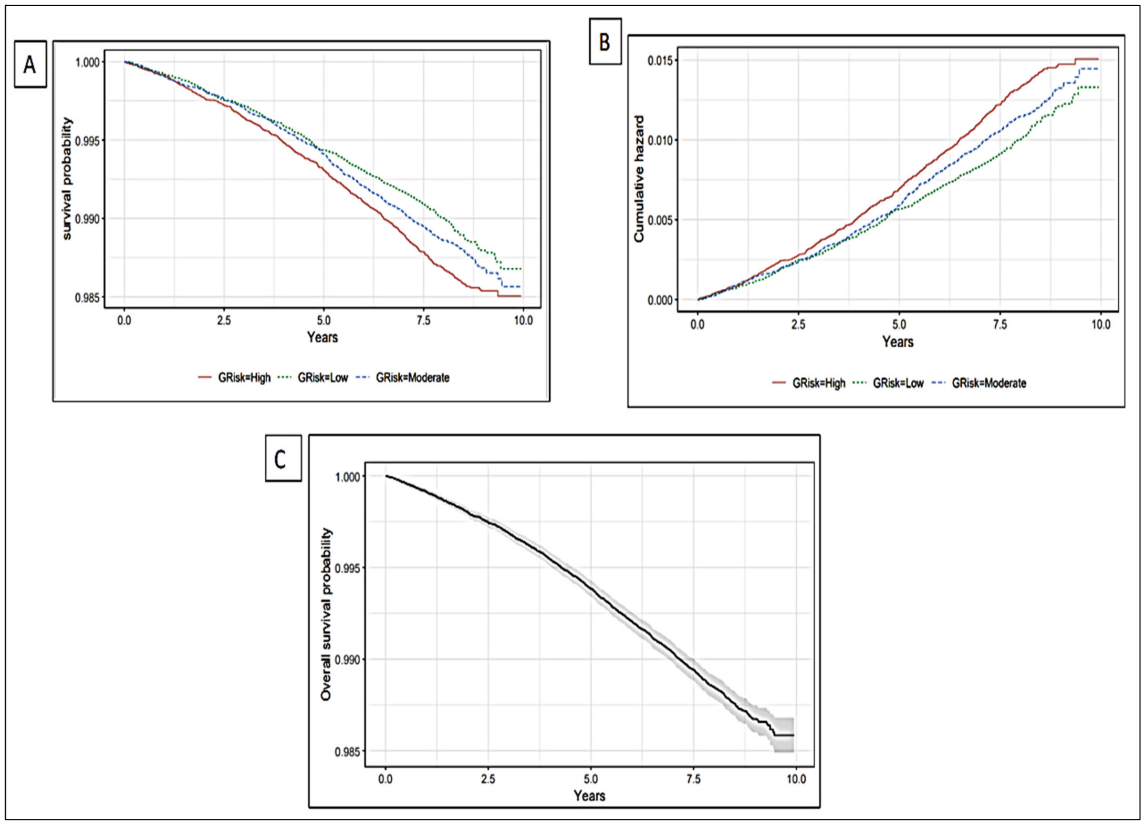


Figure 3. Survival probability and cumulative hazard plot stratified by genetic risk level: (**A**) Survival probability plot stratified by genetic risk level. (**B**) Cumulative hazard plot stratified by genetic risk level. (**C**) Overall survival probability of the study population. (**A**,**B**) demonstrate the difference in the risk of stroke between genetic liability categories. (**C**) illustrates the change in the risk of stroke over time. The grey area surrounding survival probability line in panel (**C**) represents the 95% confidence interval.

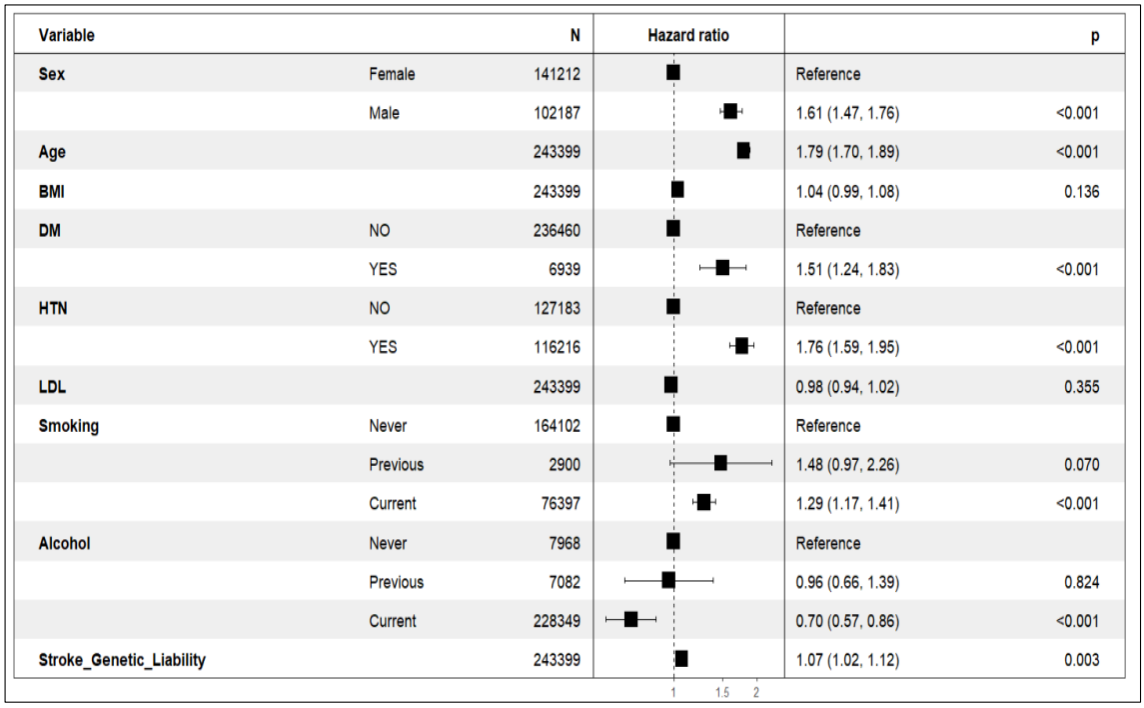


Figure 4. Forest plot of the full Cox proportional hazard model: The vertical line at the hazard ratio (HR) = 1 is the reference line. The horizontal line represents the confidence interval (CI). HTN: hypertension.

5.3. Prediction Value of the Conventional Factors

Table 4 summarizes the performance of the prediction models in the testing set. We considered predictions only up to the median follow-up time of 8.22 years.

The Cox proportional hazard model with the conventional risk factors (the model without genetic liability) showed a moderate performance and discrimination (AUC= 69.43; 95% CI = 67.30, 71.56; BS = 0.01, and ICI = 0.002) compared with the gradient boosting machines approach (AUC = 69.34; 95% Cl = 67.23, 71.50; BS = 0.01, and ICI = 0.001), the decision tree models (AUC = 67.58; 95% CI = 65.46, 69.70, BS = 0.01, and ICI = 0.001), and the random forest model, which showed the lowest performance (AUC = 65.62; 95% CI = 65.48, 67.55, BS = 0.01, and ICI = 0.003). The ROC plots of these models are presented in Supplementary Figures S5–S9 in the Supplementary Material. The result from the decision tree model indicates that age, hypertension, and sex are the most relevant predictors of stroke.

Table 3. Assessment of the proportional hazard (PH) assumption using the global Schoenfield test.

Characteristics	Chi-Square	df	*p*-Value
Sex	0.42	1	0.52
Age	0.82	1	0.37
BMI	7.49	1	0.01
DM	0.08	1	0.78
HTN	0.14	1	0.71
LDL	0.36	1	0.55
Smoking	1.40	1	0.24
Alcohol	0.12	1	0.73
SGL	2.48	1	0.12
GLOBAL	13.43	9	0.14

The table illustrates an assessment of the proportional hazard (PH) assumption using the global Schoenfield test. The test indicated a global *p*-value of 0.14, indicating no significant time-dependent joint effect on the covariates. SGL: stroke genetic liability; BMI: body mass index; LDL: low-density lipoprotein cholesterol; DM: diabetes mellitus; HTN: hypertension; df: degree of freedom.

Table 4. The result of the prediction value of the stroke genetic liability score for incident stroke in the UKB.

	Models	AUC 95%CI	NRI (95% CI)	*p*-Value for NRI	IDI (95% CI)	*p*-Value for IDI	Brier Score	ICI
Coxph	Model 1	69.43 (67.30, 71.56)	REF	REF	REF	REF	0.01	0.002
	Model 2	69.54 (67.40, 71.68)	0.20 (0.119, 0.285)	0.00	1.0×10^{-4} (0.000, 3.0×10^{-4})	0.14	0.01	0.002
GBM	Model 1	69.34 (67.23, 71.50)	REF	REF	REF	REF	0.01	0.001
	Model 2	69.38 (67.26, 71.50)	−0.11 (−0.193, −0.027)	0.01	0.00 (-1.0×10^{-4}, 1.0×10^{-4})	0.61	0.01	0.001
DT **	Model 1	61.40 (59.30, 63.40)	REF	REF	REF	REF	0.01	0.001
	Model 2	61.40 (59.30, 63.40)	0.00 (0.00, 0.00)	NaN	0.00 (0.000, 0.000)	NaN	0.01	0.001
DT	Model 1	67.58 (65.46, 69.70)	REF	REF	REF	REF	0.01	0.001
	Model 2	67.58 (65.46, 69.70)	REF	REF	REF	REF	0.01	0.001
RF	Model 1	65.62 (63.48, 67.75)	REF	REF	REF	REF	0.01	0.003
	Model 2	65.35 (63.18, 67.52)	0.17 (0.087, 0.249)	5.0×10^{-5}	0.00 (-7.0×10^{-4}, 8.0×10^{-4})	0.98	0.01	0.003

Model 1 (the basic model) features: age, sex, BMI, HTN, DM, LDL, smoking status, and alcohol status. Model 2 features: age, sex, BMI, HTN, DM, LDL, smoking status, alcohol status, and genetic liability. BMI: body mass index; LDL: low-density lipoprotein cholesterol; HTN: hypertension; DM: diabetes mellitus. DT **: decision tree built without pruning parameters. REF: reference; NaN: not a number; NRI: continuous net reclassification index; IDI: integrated discrimination; ICI: integrated calibrated index. ICI is based on a calibration curve estimated for a time-to-event outcome (time = median 8.20 years of follow-up) via a restricted cubic spline using complementary log–log transformed predicted probabilities with the "pmcalibration" function in the R program. In the reclassification analysis, the decision tree (DT) did not remarkably enhance predictions over the baseline (reference). This causes the standard error to be zero, and the NRI and IDI statistics to be near zero, resulting in NaN *p*-values.

5.4. Prediction Value of Genetic Liability

The prediction value of the Cox proportional hazards model improved slightly when the stroke genetic liability was incorporated into the model with conventional risk factors (AUC = 69.54; 95% CI = 67.40, 71.68; AUC change = 0.16%; Table 4). We also observed a slight improvement in risk reclassification, leading to an overall NRI value of 0.20 (95% CI = 0.119, 0.285; *p*-value = 0.00; Table 4). The IDI value of the Cox proportional hazard was negligibly improved by 1.0×10^{-4} (95% CI = 0.000, 3.0×10^{-4}; *p*-value = 0.14; Table 4).

The gradient boosting machine model slightly improved in prediction performance (AUC = 69.38; 95% CI = 67.26, 71.50; Table 4) but deteriorated in NRI by a value of −0.11 (95% CI = −0.193, −0.027; *p*-value = 0.01; Table 4) after adding the stroke genetic liability. There was no improvement in the overall IDI value using any of the machine learning models.

Using decision tree (AUC = 67.58, 95% Cl = 65.46, 69.70, BS = 0.01, and ICI = 0.001) or random forest (AUC = 65.35; 95% CI = 65.48, 67.55, BS = 0.01, and ICI = 0.003) models, no improvement in prediction performance was observed adding genetic liability (Table 4). The overall NRI for random forest was improved by NRI = 0.17 (95% CI = 0.087, 0.249; *p*-value = 5.0×10^{-5}; Table 4) but not for the decision tree technique. The ROC plots of these models are presented in Supplementary Figures S5–S9 in the Supplementary Material.

6. Discussion

6.1. Main Findings

The present study included genome-wide stroke genetic liability (using 252,903 genetic variants) for 243,399 participants of European descent over a median follow-up of 8.22 years. Our findings indicate that (1) the genome-wide stroke genetic liability is independently associated with the risk of stroke, (2) a prediction model integrating the genome-wide stroke genetic liability provides a slight improvement in prediction performance beyond the conventional risk factor for stroke, and (3) the Cox proportional hazard method showed better prediction performance than machine learning models (random forest, gradient boosting machines, and decision tree) with or without incorporation of genetic liability in the model.

This study's first finding, i.e., that stroke genome-wide genetic liability increases the risk of stroke, is consistent with previous studies [22–26,29] including studies by Myserlis et al. [22], Rutten-Jacobs et al. [23], Yang et al. [24], Abraham et al. [25], Verbaas et al. [26], and Hachiya [29] that reported that stroke genetic liability is a strong independent predictor of risk of future stroke occurrences.

These previous studies mainly calculated stroke genetic liability based on a limited selection of single-nucleotide polymorphisms (SNPs) that have strong associations with the traits. Our result is a step forward in the sense that we present the risk of stroke imposed by a whole-genome genetic liability of stroke in a European setting. Yang et al. [24] estimated a whole-genome genetic liability of stroke (stroke and its subtypes) in China Kadoorie Biobank and showed that the genetic liability of stroke increases risks of any stroke (14%), ischemic stroke (7%), and intracerebral hemorrhage (10%). We observed a 15% greater risk of any stroke among European participants with a high genome-wide stroke genetic liability compared with those with a low genetic liability which is comparable to the study by Yang et al. [24] in a Chinese population. Our study also differs from previous studies, including the definition or classification of the outcome and sample characteristics. We defined stroke events as any cases of (1) ischemic stroke, (2) intracerebral hemorrhage, (3) subarachnoid hemorrhage, (4) other cerebrovascular disease, or (5) stroke that is not specified as hemorrhage or infarction. Thus, we captured a broader definition of stroke, which could have increased the stroke diversity in our analysis. To investigate the relationship

between genetic liability and stroke, Rutten-Jacobs et al. [23] generated a genetic liability from 90 SNPs associated with stroke (at a *p*-value less than 1×10^{-5}). They demonstrated a 7 to 13% increase in the risk of stroke for each standard deviation increase in genetic liability. Myserlis et al. [22] and Abraham et al. [25] included the genetic liability of stroke within a meta-scoring technique that combined 19–21 distinct genetic liabilities to form a metaGRS. These studies found that the metaGRS was associated with an increased risk of incidence of intracerebral hemorrhage [22] and ischemic stroke [25]. Myserlis et al. showed a 15% increase in the risk of intracerebral hemorrhage and Abraham et al. showed a 26% increase in the risk of ischemic stroke for each standard deviation increase in the metaGRS. The association was stronger than any of the individual genetic liabilities included in the metaGRS. However, the results from Myserlis et al. and Abraham et al. did not distinguish the effect of the genetic liability of stroke per se, as the stroke genetic liability was integrated into a MetaGRS comprising 19–21 distinct genetic liabilities for various traits. Abraham et al. included several genetic liabilities for multiple stroke-related phenotypes, including ischemic stroke, any stroke, small vessel stroke, large artery stroke, cardioembolic stroke, and several stroke risk factors in their metaGRS. Myserlis et al. included genetic liabilities for multiple phonotypes including white matter hemorrhage (n = 87,951 SNPs) and small vessel stroke (n = 2162 SNPs) within the metaGRS calculation. While metaGRS has been found to improve risk prediction, there may be some biases in prediction performance because it was built using elastic-net regression. Additionally, certain SNPs included in the calculation of individual phenotypes' genetic liabilities may be associated with several phenotypes [26]. Therefore, the metaGRS may contain overlapping information due to possible correlation among the genetic liabilities included in the metaGRS [69]. Our approach to considering genome-wide genetic liability for stroke aimed to capture the polygenic component of stroke, i.e., we had no statistical significance threshold for the selection of SNPs associated with stroke. Thus, we included all SNPs, even those with small or non-significant effects. It is known that this approach would increase the accuracy of the effect estimated for genetic liability and would, therefore, improve accuracy in the identification of high-risk individuals [70,71].

Our prediction models demonstrated that the genome-wide stroke genetic liability may slightly enhance (1) overall stroke prediction performance to distinguish the cases (the Cox proportional hazards model and the gradient boosting machine) and (2) correct classification of individuals at risk beyond conventional risk factors (the Cox proportional hazards model and random forest). However, none of our models demonstrated statistically improved predicted probabilities for cases and non-cases based on IDI.

Our findings from prediction analysis are supported by the results reported in previous studies [40,72,73] which observed that including both genetic liability and conventional risk factors in risk prediction models improves the discrimination performance compared to using only conventional risk factors. Papadopoulou et al. [40] used genetic liability based on 28 SNPs in a European population focused on ischemic stroke in patients with atrial fibrillation (AF). They observed that XGBoost performed better than the CHA2DS2-VASc model, an existing clinical model for calculating stroke risk for patients with atrial fibrillation. Cárcel-Márquez et al. [72] used genetic liability based on 93 SNPs to predict cardioembolic stroke in the European population using logistic regression while Jung et al. [73] used genetic liability based on 16 SNPS to predict stroke in a Korean population using Cox proportional hazard regression. Our best model performed better than the models of Papadopoulou et al. [40] and Jung et al. [73]. Our study population differed from the populations studied by Papadopoulou et al. [40] and Jung et al. [73]. However, the risk values identified in the current study are smaller than those published by

Cárcel-Márquez et al. [72]. It should be emphasized that Cárcel-Márquez and colleagues did not use time-to-event data and instead focused on cardioembolic stroke.

Unlike previous studies, which implied that machine learning algorithms outperform traditional statistical approaches in the prediction of stroke [32,41,42], the current study indicated that the Cox proportional hazard regression models outperformed all the machine learning models in the context of time-to-event data for stroke. This could be due to the small number of events (1973 stroke events) and few predictors (up to 9 predictors) in this study. These two reasons are considered as reasons for Cox models to outperform machine learning [74]. We found that genetic liability improved stroke risk classification for less than 1% of the subjects. Health economy studies could consider investigating if using this information in the identification of high-risk individuals to target for stroke prevention programs could make a significant cost-effective change in stroke-related expenses.

The large sample size of UK Biobank and the number of incident strokes enabled the statistical power for our analysis in which we used time-to-event data for over 200,000 individuals of European ancestry, with a median follow-up of 8.22 years. A distinctive feature and the strength of our study compared with previous studies is that we generated genetic liability for stroke using over 250,000 genetic variants.

Validation in external cohort datasets could improve the precision of our findings. To minimize lack of validation in external cohorts, we internally validated our machine learning models in the testing set, where we randomly partitioned the data into a training set (70% of participants) for developing the prediction models and a testing set (30% of participants) to evaluate the prediction models' performance.

6.2. Implication

Genetic predisposition to stroke had minimal impact on improving stroke risk prediction, benefiting approximately one percent of the population. Since genetic liability improved prediction for only a small percentage of the population, its application in clinical practice is uncertain. Conventional risk factors may still have more influence on the prediction of stroke. The findings suggest that genetic liability alone has limited predictive value for most people, but they might still have a role in highly targeted interventions.

In terms of cost effectiveness, given that only a small percentage of the population benefits from genetic risk scores, health economics studies are needed to establish if the costs of genetic testing outweigh the potential improvements in stroke prevention.

7. Conclusions

In conclusion, incorporating genetic liability into stroke risk prediction models could slightly improve prediction performance and should be considered when predicting the risk of stroke. Cox proportional hazard models should be given priority over machine learning models in the prediction of the risk of stroke.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/healthcare13091003/s1, Supplementary Data S1—Supplementary Data S1: List of genetic variants summary statistics used to construct the genetic risk scores. Table S1: Overview of previous studies investigating the effect of genetic liability on risk of stroke. Figure S1: Overview of the process to create genetic liability for stroke within the UK Biobank. Figure S2: Workflow diagram illustrating the inputs and output for machine learning models. Figure S3: Overview of the modeling and predictions process for using machine learning and genome-wide genetic liability for prediction of risk of stroke within the UK Biobank. Figure S4: Schoenfield test results of full Cox proportional hazard model. Figure S5: Roc plot of coxph models. Figure S6: Roc plot of Gradient boosting models. Figure S7: Roc plot of decision tree models (using pruning). Figure S8: Roc plot of decision tree models (without pruning). Figure S9: Roc plot of Random Forest models.

Author Contributions: Conceptualization, R.P.; Data curation, Formal analysis, G.M.; Investigation, G.M.; Methodology, G.M. and R.P.; Project administration, G.M. and R.P.; Resources, R.P.; Supervision, R.P.; Writing—original draft, G.M.; Writing—review and editing, G.M. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (Ethics Committee) of Brunel University of London, College of Health, Medicine, and Life Sciences (27684-LR-Jan/2021-29901-1 on 5 February 2021).

Informed Consent Statement: Informed consent was obtained from all participants participated in the UK Biobank.

Data Availability Statement: The data used in this study is available on request from the UK Biobank.

Acknowledgments: This research was conducted using the UK Biobank under Application Number 60549 (www.ukbiobank.ac.uk (accessed on 5 February 2021)). The UK Biobank is generously supported by its founding funders, the Wellcome Trust and the UK Medical Research Council, as well as by the British Heart Foundation, Cancer Research UK, the Department of Health, the Northwest Regional Development Agency, and the Scottish Government. The MEGASTROKE project received funding from sources specified at https://megastroke.org/acknowledgements.html (accessed on 13 September 2022).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Roth, G.A.; Johnson, C.; Abajobir, A.; Abd-Allah, F.; Abera, S.F.; Abyu, G.; Ahmed, M.; Aksut, B.; Alam, T.; Alam, K.; et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **2017**, *70*, 1–25. [CrossRef] [PubMed]
2. Krishnamurthi, R.; Ikeda, T.; Feigin, V. 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*, 171–179. [CrossRef]
3. Dhamoon, M.S.; Tai, W.; Boden-Albala, B.; Rundek, T.; Paik, M.C.; Sacco, R.L.; Elkind, M.S.V. Risk of Myocardial Infarction or Vascular Death After First Ischemic Stroke: The Northern Manhattan Study. *Stroke* **2007**, *38*, 1752–1758. [CrossRef] [PubMed]
4. Dhamoon, M.S.; Sciacca, R.R.; Rundek, T.; Sacco, R.L.; Elkind, M.S.V. Recurrent stroke and cardiac risks after first ischemic stroke: The Northern Manhattan study. *Neurology* **2006**, *66*, 641–646. [CrossRef] [PubMed]
5. Kyme, C. After ischemic stroke, patients are at higher risk of recurrent stroke than of cardiac events. *Nat. Clin. Pract. Cardiovasc. Med.* **2005**, *2*, 436. [CrossRef]
6. Engstad, T.; Viitanen, M.; Arnesen, E. Predictors of Death Among Long-Term Stroke Survivors. *Stroke* **2003**, *34*, 2876–2880. [CrossRef]
7. 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*, 277–282. [CrossRef]
8. Boehme, A.K.; Esenwa, C.; Elkind, M.S.V. Stroke Risk Factors, Genetics, and Prevention. *Circ. Res.* **2017**, *120*, 472–495. [CrossRef]
9. Benjamin, E.J.; Blaha, M.J.; Chiuve, S.E.; Cushman, M.; Das, S.R.; Deo, R.; de Ferranti, S.D.; Floyd, J.; Fornage, M.; Gillespie, C.; et al. Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association. *Circulation* **2017**, *135*, e146–e603. [CrossRef]
10. Bak, S.; Gaist, D.; Sindrup, S.H.; Skytthe, A.; Christensen, K. Genetic Liability in Stroke: A Long-Term Follow-Up Study of Danish Twins. *Stroke* **2002**, *33*, 769–774. [CrossRef]
11. Malik, R.; Chauhan, G.; Traylor, M.; Okada, Y.; Giese, A.K.; Laan, S.; Chong, M.; Adams, H.; Ago, T.; Almgren, P.; et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **2018**, *50*, 524–537. [CrossRef]
12. Malik, R.; Rannikmäe, K.; Traylor, M.; Georgakis, M.K.; Sargurupremraj, M.; Markus, H.S.; Hopewell, J.C.; Debette, S.; Sudlow, C.L.M.; Dichgans, M. Genome-wide meta-analysis identifies 3 novel loci associated with stroke. *Ann. Neurol.* **2018**, *84*, 934–939. [CrossRef]
13. American Heart Association News. New tool brings big changes to cardiovascular disease predictions. In *Premium Official News*; American Heart Association News: Dallas, TX, USA, 2023.

14. Elias, M.F.; Sullivan, L.M.; D'Agostino, R.B.; Elias, P.K.; Beiser, A.; Au, R.; Seshadri, S.; DeCarli, C.; Wolf, P.A. Framingham Stroke Risk Profile and Lowered Cognitive Performance. *Stroke* **2004**, *35*, 404–409. [CrossRef]
15. Bohr, A.; Memarzadeh, K. *The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare*; Academic Press: Cambridge, MA, USA, 2020; pp. 25–60.
16. Knowles, J.W.; Ashley, E.A. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med.* **2018**, *15*, e1002546. [CrossRef] [PubMed]
17. Traylor, M.; Farrall, M.; Holliday, E.G.; Sudlow, C.; Hopewell, J.C.; Cheng, Y.C.; Fornage, M.; Ikram, M.A.; Malik, R.; Bevan, S.; et al. Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE Collaboration): A meta-analysis of genome-wide association studies. *Lancet Neurol.* **2012**, *11*, 951–962. [CrossRef] [PubMed]
18. Abraham, G.; Rutten-Jacobs, L.; Inouye, M. Risk Prediction Using Polygenic Risk Scores for Prevention of Stroke and Other Cardiovascular Diseases. *Stroke* **2021**, *52*, 2983–2991. [CrossRef] [PubMed]
19. Gschwendtner, A.; Dichgans, M. Genetics of ischemic stroke. *Nervenarzt* **2013**, *84*, 166. [CrossRef]
20. Della-Morte, D.; Guadagni, F.; Palmirotta, R.; Testa, G.; Caso, V.; Paciaroni, M.; Abete, P.; Rengo, F.; Ferroni, P.; Sacco, R.L.; et al. Genetics of ischemic stroke, stroke-related risk factors, stroke precursors and treatments. *Pharmacogenomics* **2012**, *13*, 595–613. [CrossRef]
21. Mishra, A.; Malik, R.; Hachiya, T.; Jürgenson, T.; Namba, S.; Posner, D.C.; Kamanu, F.K.; Koido, M.; Le Grand, Q.; Shi, M.; et al. Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* **2022**, *611*, 115–123. [CrossRef]
22. Myserlis, E.P.; Georgakis, M.K.; Demel, S.L.; Sekar, P.; Chung, J.; Malik, R.; Hyacinth, H.I.; Comeau, M.E.; Falcone, G.J.; Langefeld, C.D.; et al. A Genomic Risk Score Identifies Individuals at High Risk for Intracerebral Hemorrhage. *Stroke* **2023**, *54*, 973–982. [CrossRef]
23. Rutten-Jacobs, L.C.; Larsson, S.C.; Malik, R.; Rannikmäe, K.; Sudlow, C.L.; Dichgans, M.; Markus, H.S.; Traylor, M. Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: Cohort study of 306 473 UK Biobank participants. *BMJ* **2018**, *363*, k4168. [CrossRef] [PubMed]
24. Yang, S.; Sun, Z.; Sun, D.; Yu, C.; Guo, Y.; Sun, D.; Pang, Y.; Pei, P.; Yang, L.; Millwood, I.Y.; et al. Associations of polygenic risk scores with risks of stroke and its subtypes in Chinese. *Stroke Vasc. Neurol.* **2024**, *9*, 399–406. [CrossRef] [PubMed]
25. Abraham, G.; Malik, R.; Yonova-Doing, E.; Salim, A.; Wang, T.; Danesh, J.; Butterworth, A.S.; Howson, J.M.M.; Inouye, M.; Dichgans, M. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat. Commun.* **2019**, *10*, 5819. [CrossRef]
26. Verbaas, C.; Fornage, M.; Bis, J.C.; Choi, S.H.; Psaty, B.M.; Meigs, J.B.; Rao, M.; Nalls, M.; Fontes, J.D.; O'Donnell, C.J.; et al. Predicting Stroke Through Genetic Risk Functions the CHARGE Risk Score Project. *Stroke* **2014**, *45*, 403–412. [CrossRef]
27. Bakker, M.K.; Kanning, J.P.; Abraham, G.; Martinsen, A.E.; Winsvold, B.S.; Zwart, J.A.; Bourcier, R.; Sawada, T.; Koido, M.; Kamatani, Y.; et al. Genetic Risk Score for Intracranial Aneurysms: Prediction of Subarachnoid Hemorrhage and Role in Clinical Heterogeneity. *Stroke* **2023**, *54*, 810–818. [CrossRef] [PubMed]
28. Malik, R.; Bevan, S.; Nalls, M.A.; Holliday, E.G.; Devan, W.J.; Cheng, Y.C.; Ibrahim-Verbaas, C.A.; Verhaaren, B.F.; Bis, J.C.; Joon, A.Y.; et al. Multilocus Genetic Risk Score Associates with Ischemic Stroke in Case–Control and Prospective Cohort Studies. *Stroke* **2014**, *45*, 394–402. [CrossRef]
29. Hachiya, T.; Hata, J.; Hirakawa, Y.; Yoshida, D.; Furuta, Y.; Kitazono, T.; Shimizu, A.; Ninomiya, T. Genome-Wide Polygenic Score and the Risk of Ischemic Stroke in a Prospective Cohort: The Hisayama Study. *Stroke* **2020**, *51*, 759–765. [CrossRef]
30. Hachiya, T.; Kamatani, Y.; Takahashi, A.; Hata, J.; Furukawa, R.; Shiwa, Y.; Yamaji, T.; Hara, M.; Tanno, K.; Ohmomo, H.; et al. Genetic Predisposition to Ischemic Stroke: A Polygenic Risk Score. *Stroke* **2017**, *48*, 253–258. [CrossRef]
31. Lynch, C.M.; Abdollahi, B.; Fuqua, J.D.; de Carlo, A.R.; Bartholomai, J.A.; Balgemann, R.N.; van Berkel, V.H.; Frieboes, H.B. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int. J. Med. Inform.* **2017**, *108*, 1–8. [CrossRef]
32. Chun, M.; Clarke, R.; Cairns, B.J.; Clifton, D.; Bennett, D.; Chen, Y.; Guo, Y.; Pei, P.; Lv, J.; Yu, C.; et al. Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1719–1727. [CrossRef]
33. MacCarthy, G.; Pazoki, R. Using Machine Learning to Evaluate the Value of Genetic Liabilities in the Classification of Hypertension within the UK Biobank. *J. Clin. Med.* **2024**, *13*, 2955. [CrossRef] [PubMed]
34. Schjerven, F.E.; Ingeström, E.M.L.; Steinsland, I.; Lindseth, F. Development of risk models of incident hypertension using machine learning on the HUNT study data. *Sci. Rep.* **2024**, *14*, 5609. [CrossRef] [PubMed]
35. Wongvibulsin, S.; Wu, K.C.; Zeger, S.L. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med. Res. Methodol.* **2019**, *20*, 1. [CrossRef] [PubMed]
36. Wang, Y.; Zhang, L.; Niu, M.; Li, R.; Tu, R.; Liu, X.; Hou, J.; Mao, Z.; Wang, Z.; Wang, C. Genetic Risk Score Increased Discriminant Efficiency of Predictive Models for Type 2 Diabetes Mellitus Using Machine Learning: Cohort Study. *Front. Public Health* **2021**, *9*, 606711. [CrossRef]

37. Datema, F.R.; Moya, A.; Krause, P.; Bäck, T.; Willmes, L.; Langeveld, T.; Baatenburg de Jong, R.J.; Blom, H.M. Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. *Head Neck* **2012**, *34*, 50–58. [CrossRef]
38. Qiu, X.; Gao, J.; Yang, J.; Hu, J.; Hu, W.; Kong, L.; Lu, J.J. A Comparison Study of Machine Learning (Random Survival Forest) and Classic Statistic (Cox Proportional Hazards) for Predicting Progression in High-Grade Glioma after Proton and Carbon Ion Radiotherapy. *Front. Oncol.* **2020**, *10*, 551420. [CrossRef]
39. Xu, L.; Cai, L.; Zhu, Z.; Chen, G. Comparison of the cox regression to machine learning in predicting the survival of anaplastic thyroid carcinoma. *BMC Endocr. Disord.* **2023**, *23*, 129. [CrossRef]
40. Papadopoulou, A.; Harding, D.; Slabaugh, G.; Marouli, E.; Deloukas, P. Prediction of atrial fibrillation and stroke using machine learning models in UK Biobank. *Heliyon* **2024**, *10*, e28034. [CrossRef]
41. Wang, Y.; Deng, Y.; Tan, Y.; Zhou, M.; Jiang, Y.; Liu, B. A comparison of random survival forest and Cox regression for prediction of mortality in patients with hemorrhagic stroke. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 215. [CrossRef]
42. Chen, Y.; Chung, J.; Yeh, Y.; Lou, S.; Lin, H.; Lin, C.; Hsien, H.; Hung, K.; Yeh, S.J.; Shi, H. Predicting 30-Day Readmission for Stroke Using Machine Learning Algorithms: A Prospective Cohort Study. *Front. Neurol.* **2022**, *13*, 875491. [CrossRef]
43. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **2015**, *12*, e1001779. [CrossRef]
44. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'connell, J.; et al. Genome-wide genetic data on ~500,000 UK biobank participants. *bioRxiv* **2017**. [CrossRef]
45. Welsh, S.; Peakman, T.; Sheard, S.; Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genom.* **2017**, *18*, 26. [CrossRef] [PubMed]
46. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; et al. The Uk biobank resource with deep phenotyping and genomic data. *Nature* **2018**, *562*, 203–209. [CrossRef] [PubMed]
47. Marchini, J.; O'Connell, J.; Delaneau, O.; Sharp, K.; Kretzschmar, W.; Band, G.; McCarthy, S.; Petkova, D.; Bycroft, C.; Freeman, C.; et al. UK Biobank Phasing and Imputation Documentation Contributors to UK Biobank Phasing and Imputation. 2015. Available online: https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/impute_ukb_v1.pdf (accessed on 1 December 2023).
48. Sacks, D.B.; Arnold, M.; Bakris, G.L.; Bruns, D.E.; Horvath, A.R.; Kirkman, M.S.; Lernmark, A.; Metzger, B.E.; Nathan, D.M. Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. *Clin. Chem.* **2011**, *57*, e1–e47. [CrossRef]
49. Flack, J.M.; Adekola, B. Blood pressure and the new ACC/AHA hypertension guidelines. *Trends Cardiovasc. Med.* **2020**, *30*, 160–164. [CrossRef]
50. Chobanian, A.V.; Bakris, G.L.; Black, H.R.; Cushman, W.C.; Green, L.A.; Izzo, J.; Joseph, L.; Jones, D.W.; Materson, B.J.; Oparil, S.; et al. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: The JNC 7 Report. *J. Am. Med. Assoc.* **2003**, *289*, 2560–2571. [CrossRef]
51. Pazoki, R.; Dehghan, A.; Evangelou, E.; Warren, H.; Gao, H.; Caulfield, M.; Elliott, P.; Tzoulaki, I. Genetic Predisposition to High Blood Pressure and Lifestyle Factors: Associations with Midlife Blood Pressure Levels and Cardiovascular Events. *Circulation* **2018**, *137*, 653–661. [CrossRef]
52. Marees, A.T.; de Kluiver, H.; Stringer, S.; Vorspan, F.; Curis, E.; Marie-Claire, C.; Derks, E.M. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **2018**, *27*, e1608. [CrossRef]
53. Chang, C.C. Data management and summary statistics with PLINK. *Methods Mol. Biol.* **2020**, *2090*, 49–65.
54. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef] [PubMed]
55. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2010.
56. Tabachnick, B.G.; Fidell, L.S. *Using Multivariate Statistics*, 6th ed.; Pearson: Boston, MA, USA, 2013.
57. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J.R.G.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. [CrossRef]
58. Deo, S.V.; Deo, V.; Sundaram, V. Survival analysis—Part 2: Cox proportional hazards model. *Indian J. Thorac. Cardiovasc. Surg.* **2021**, *37*, 229–233. [CrossRef] [PubMed]
59. Abd ElHafeez, S.; D'Arrigo, G.; Leonardis, D.; Fusaro, M.; Tripepi, G.; Roumeliotis, S. Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. *Oxidative Med. Cell. Longev.* **2021**, *2021*, 1302811. [CrossRef]
60. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning*, 2nd ed.; corrected at 5 print ed.; Springer: New York, NY, USA, 2011.

61. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
62. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Largin Margin Classif.* **1999**, *10*, 61–74.
63. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 621–633. [CrossRef]
64. Van Calster, B.; Nieboer, D.; Vergouwe, Y.; De Cock, B.; Pencina, M.J.; Steyerberg, E.W. A calibration hierarchy for risk models was defined: From utopia to empirical data. *J. Clin. Epidemiol.* **2016**, *74*, 167–176. [CrossRef]
65. Miller, T.D.; Askew, J.W. Net reclassification improvement and integrated discrimination improvement: New standards for evaluating the incremental value of stress imaging for risk assessment. Circulation. *Cardiovasc. Imaging* **2013**, *6*, 496–498. [CrossRef]
66. McKearnan, S.B.; Wolfson, J.; Vock, D.M.; Vazquez-Benitez, G.; O'Connor, P.J. Performance of the Net Reclassification Improvement for Nonnested Models and a Novel Percentile-Based Alternative. *Am. J. Epidemiol.* **2018**, *187*, 1327–1335. [CrossRef]
67. Pencina, M.J.; D'Agostino, R.B., Sr.; Steyerberg, E.W. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* **2011**, *30*, 11–21. [CrossRef]
68. Steyerberg, E.; Vickers, A.; Cook, N.; Gerds, T.; Gonen, M.; Obuchowski, N.; Pencina, M.; Kattan, M. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* **2010**, *21*, 128–138. [CrossRef]
69. Clark, K.; Fu, W.; Liu, C.; Ho, P.; Wang, H.; Lee, W.; Chou, S.; Wang, L.; Tzeng, J. The prediction of Alzheimer's disease through multi-trait genetic modeling. *Front. Aging Neurosci.* **2023**, *15*, 1168638. [CrossRef] [PubMed]
70. Wray, N.R.; Goddard, M.E. Multi-locus models of genetic risk of disease. *Genome Med.* **2010**, *2*, 10. [CrossRef] [PubMed]
71. Wang, Y.; Namba, S.; Lopera, E.; Kerminen, S.; Tsuo, K.; Läll, K.; Kanai, M.; Zhou, W.; Favé, M.-J.; Bhatta, L.; et al. Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genom.* **2023**, *3*, 100241. [CrossRef] [PubMed]
72. Cárcel-Márquez, J.; Muiño, E.; Gallego-Fabrega, C.; Cullell, N.; Lledós, M.; Llucià-Carol, L.; Sobrino, T.; Campos, F.; Castillo, J.; Freijo, M.; et al. A Polygenic Risk Score Based on a Cardioembolic Stroke Multitrait Analysis Improves a Clinical Prediction Model for This Stroke Subtype. *Front. Cardiovasc. Med.* **2022**, *9*, 940696. [CrossRef]
73. Jung, K.J.; Hwang, S.; Lee, S.; Kim, H.C.; Jee, S.H. Traditional and Genetic Risk Score and Stroke Risk Prediction in Korea. *Korean Circ. J.* **2018**, *48*, 731–740. [CrossRef]
74. Du, M.; Haag, D.G.; Lynch, J.W.; Mittinty, M.N. Comparison of the Tree-Based Machine Learning Algorithms to Cox Regression in Predicting the Survival of Oral and Pharyngeal Cancers: Analyses Based on SEER Database. *Cancers* **2020**, *12*, 2802. [CrossRef]

Table 4 Corrected

Model Type	Model	AUC (95% CI)	NRI (95% CI)	p-value (NRI)	IDI (95% CI)	p-value (IDI)	Brier Score	ICI
CoxPH	Model 1	69.43 (67.30, 71.56)	REF	REF	REF	REF	0.01	0.002
	Model 2	69.54 (67.40, 71.68)	0.20 (0.119, 0.285)	<0.0001	1.0×10^{-4} ($0.000, 3.0\times10^{-4}$)	0.14	0.01	0.002
GBM	Model 1	69.34 (67.23, 71.50)	REF	REF	REF	REF	0.01	0.001
	Model 2	69.38 (67.26, 71.50)	-0.11 (-0.193, -0.027)	0.01	0.00 (-1.0×10^{-4},1.0×10^{-4})	0.61	0.01	0.001
DT	Model 1	61.40 (59.30, 63.40)	REF	REF	REF	REF	0.01	0.001
	Model 2	61.40 (59.30, 63.40)	0.00 (0.00, 0.00)	NaN	0.00 (0.000,0.000)	NaN	0.01	0.001
DT**	Model 1	67.58 (65.46, 69.70)	REF	REF	REF	REF	0.01	0.001
	Model 2	67.58 (65.46, 69.70)	0.00 (0.00, 0.00)	NaN	0.00 (0.000, 0.000)	NaN	0.01	0.001
RF	Model 1	65.62 (63.48, 67.75)	REF	REF	REF	REF	0.01	0.003
	Model 2	65.35 (63.18, 67.52)	0.17 (0.087, 0.249)	5.0×10^{-5}	0.00 (-7.0×10^{-4} ,8.0×10^{-5})	0.98	0.01	0.003

AUC = Area Under the Curve; **CI** = Confidence Interval; **NRI** = Net Reclassification Improvement; IDI = Integrated Discrimination Improvement; ICI = Integrated Calibration Index; **REF** = reference model; **NaN** indicates Not a Number. **CoxPH** = Cox proportional hazards model; **GBM** = Gradient Boosting Machine; **DT** = Decision Tree; **DT**** =Tuned Decision Tree; **RF** = Random Forest.

Supplementary Material for Chapter 4

Evaluation of Machine Learning and Traditional Statistical Models to Assess the Value of Stroke Genetic Liability for Prediction of Risk of Stroke Within the UK Biobank

The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/healthcare13091003/s1, Supplementary Data S1—Supplementary Data S1: List of genetic variants summary statistics used to construct the genetic risk scores. Table S1: Overview of previous studies investigating the effect of genetic liability on risk of stroke. Figure S1: Overview of the process to create genetic liability for stroke within the UK Biobank. Figure S2: Workflow diagram illustrating the inputs and output for machine learning models. Figure S3: Overview of the modeling and predictions process for using machine learning and genome-wide genetic liability for prediction of risk of stroke within the UK Biobank. Figure S4: Schoenfield test results of full Cox proportional hazard model. Figure S5: Roc plot of coxph models. Figure S6: Roc plot of Gradient boosting models. Figure S7: Roc plot of decision tree models (using pruning). Figure S8: Roc plot of decision tree models (without pruning). Figure S9: Roc plot of Random Forest models.

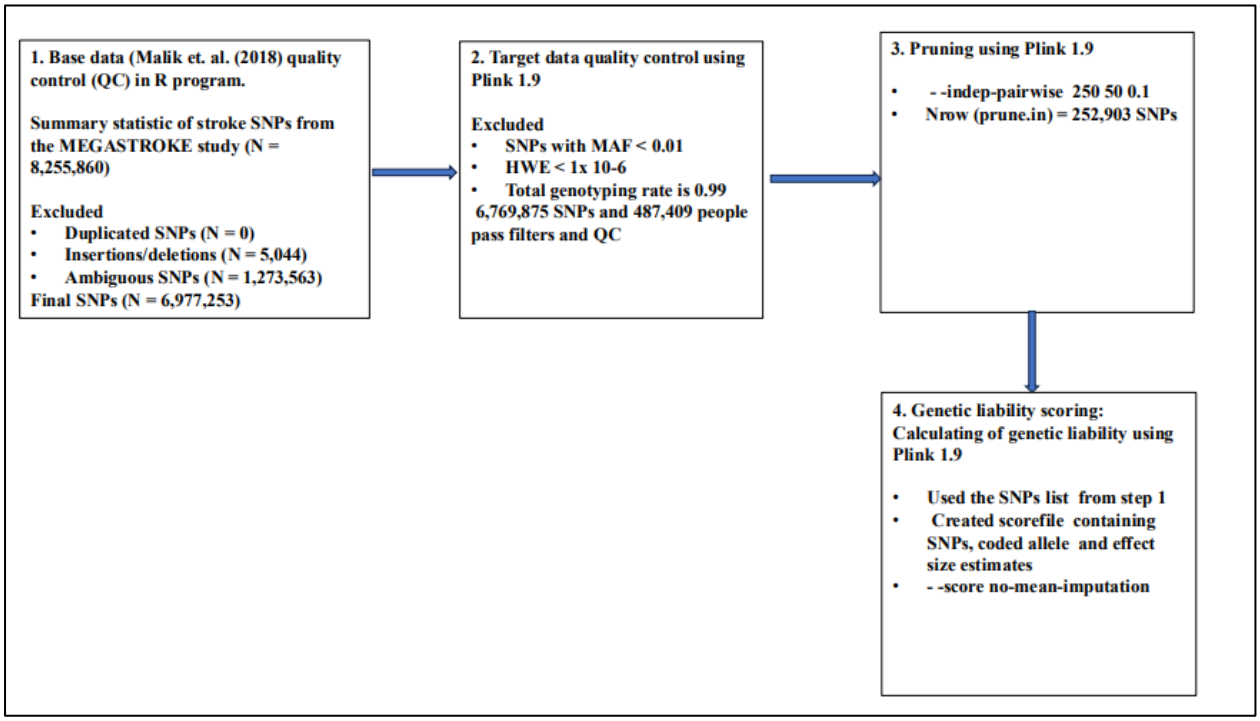
Supplementary Table 1: Overview of previous studies investigating the effect of genetic liability on risk of stroke.

Author, in-text citation	Genetic liability characteristics	Stroke outcome	Mode name	Measure of risk	Prediction performance	Ethnicity
Myserlis et al., 2023 [22]	metaGRS (Combined 21 GRS)	hemorrhage	Coxph	HR = 1.15	C-index Increased from 68.9% to 69.5% when metaGRS was added to the model.	European ancestry
Rutten-Jacobs et al., 2018 [23]	90-SNPs GRS for any stroke	All Stroke	Coxph	HR = 1.35	N/A	European ancestry
Abraham et al., 2019 [25]	metaGRS (Combined 19 GRSs)	Ischemic	Coxph	HR = 1.26	N/A	European ancestry
Papadopoulo et al., 2024[40]	28-SNPs GRS and conventional risk factors	Ischemic	ML	NA	XGBoost with ROC of 63.1%.	European Ancestry with AF
Wang et al., 2023[41]	Selected conventional risk factors	Hemorrhage	Coxph and ML	NA	RF (AUC = 87.5%) vs Coxph(AUC = 76.1%)	United states
Cárcel-Márquez et al., 2022 [73]	93-SNPs GRS and Selected conventional risk factors	cardioembolic	MTAG	NA	AUC=94.7% without GRS AUC = 95.0% when GRS was added.	European ancestry
Jung et al., 2018 [74]	16-SNPs GRS and selected	stroke	Coxph	NA	AUC= 67%	Korean

	conventional risk factors.					

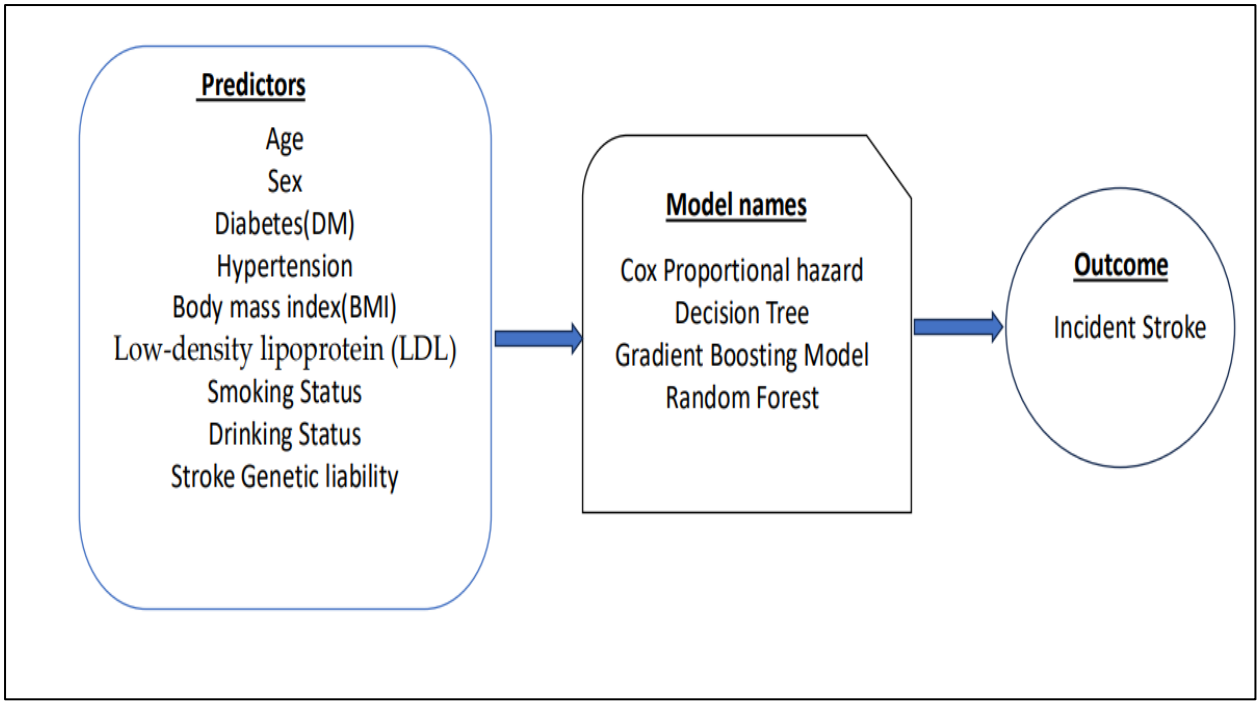
- ***MTAG***:*Multitrait Analysis of Genome Wide Association Study*

Supplementary Figure 1: Overview of the process to create genetic liability for stroke within the UK Biobank.

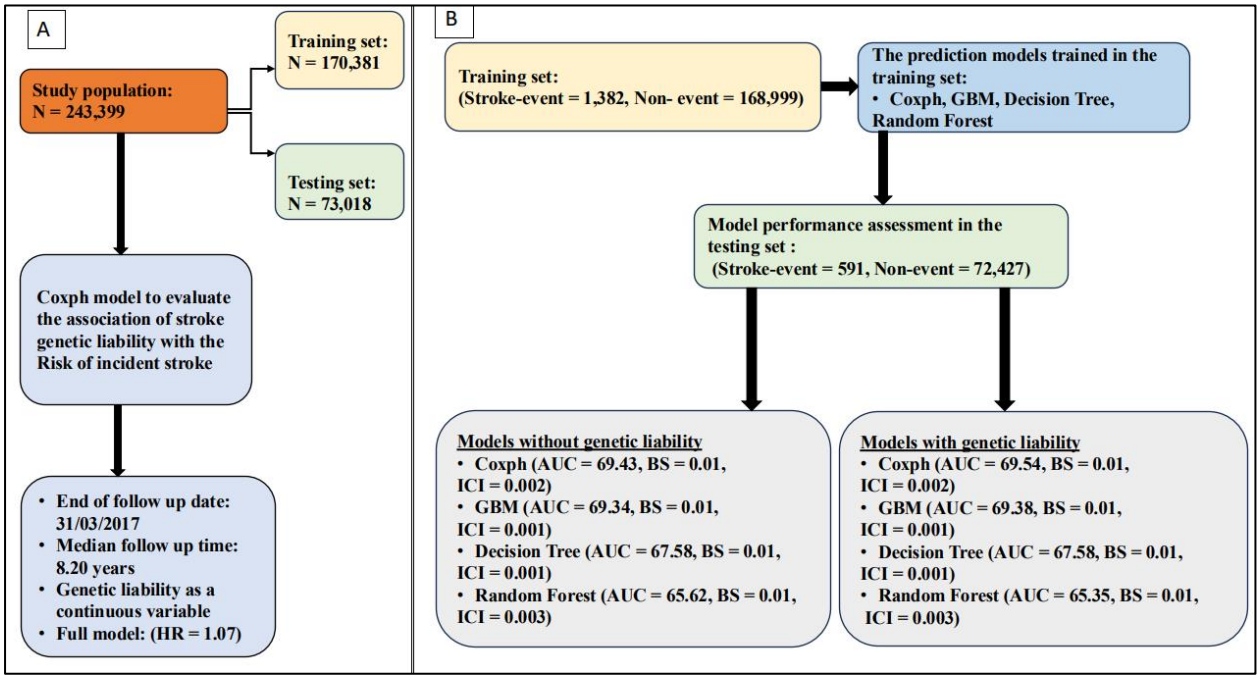


Genetic liability calculation process: SNP: Single nucleotide polymorphism; **MAF**: Minor allele frequency; **LD**: Linkage disequilibrium; **HWE**: Hardy-Weinberg Equilibrium. The SNPs were pruned with plink command - -indep-pairwise window size = 250, step size = 50, $r^2 = 0.1$. **Base data**: SNP list from Malik et. al. (2018), **Target data**: genotype data in plink binary format.

Supplementary Figure 2: Workflow diagram illustrating the inputs and output for machine learning **models.**

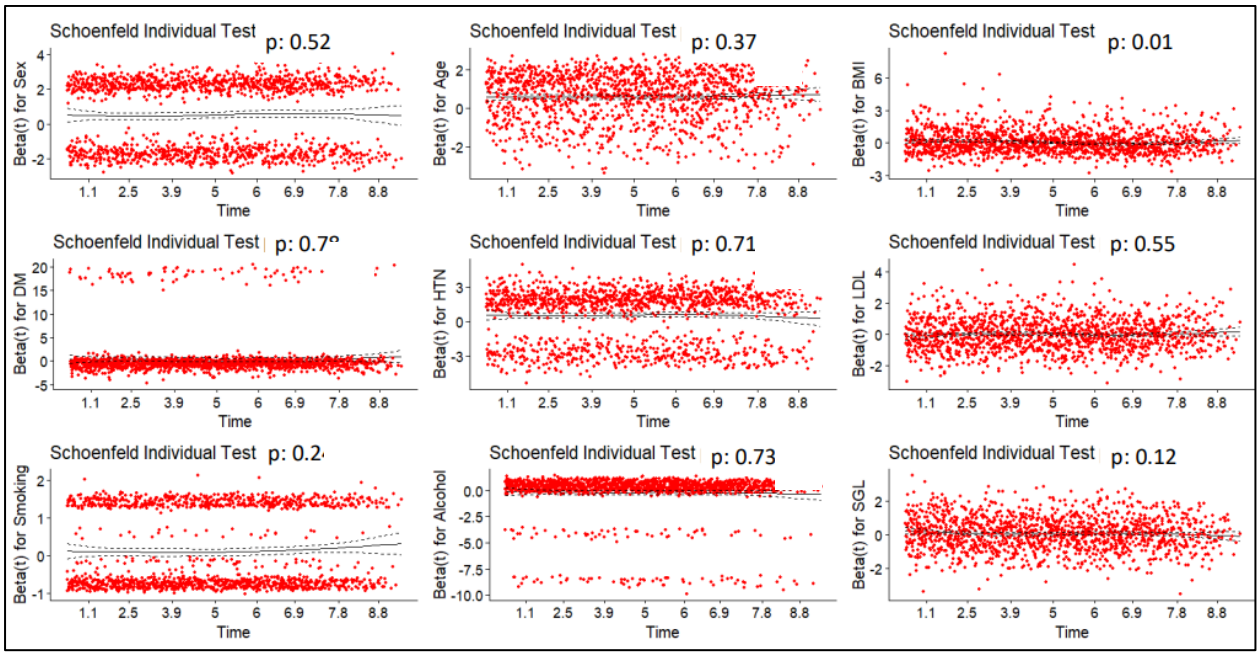


Supplementary Figure 3: Overview of the modelling and prediction process using machine learning and genome-wide genetic liability to predict stroke risk within the UK Biobank.



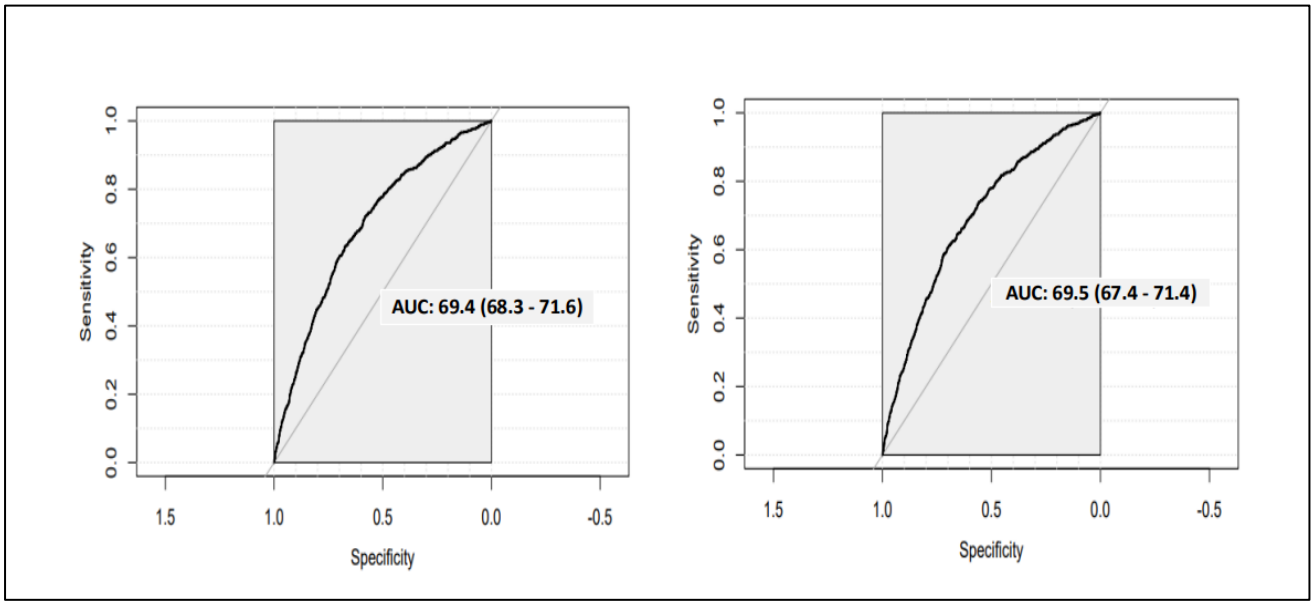
Stroke risk prediction model creation and performance evaluation: Coxph: Cox proportional hazard, **GBM**: Gradient Boosting model, **HR**= Hazard ratio, **AUC**: Area under the curve, **BS**: Brier score, ICI: Integrated calibrated index. **Panel A**: Assessing the association between genetic liability and incident stroke. **Panel B**: Stroke risk prediction modeling and performance evaluation.

Supplementary Figure 4: Schoenfeld test results of full Cox proportional hazard model.



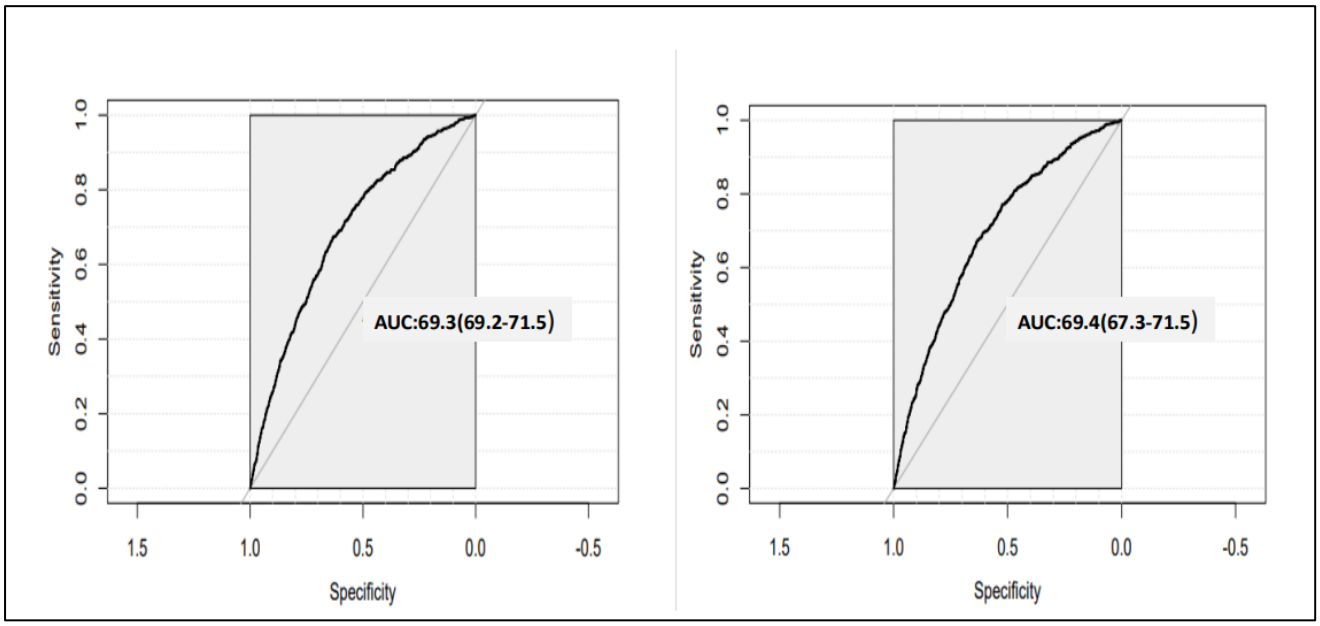
The figure illustrates an assessment of the proportional hazard (PH) assumption using the global Schoenfield test that assesses the proportional hazard assumption for all covariates from a multivariate model. The test indicated a p-value of 0.14, indicating no significant time-dependent effect on the covariates jointly. If P-value > 0.05, the test fails to reject the null hypothesis, i.e. the PH assumption would hold for the overall model, and covariates have a consistent effect over time. The individual Schoenfield test for BMI (p-value = 0.01) indicates that it does not have a consistent effect over time. Thus, BMI is adjusted within all the Cox models in the study. **BMI**: Body Mass Index, **LDL**: Low-density lipoprotein cholesterol, **SGL**: Stroke genetic liability

Supplementary Figure 5: Roc plot of Coxph models



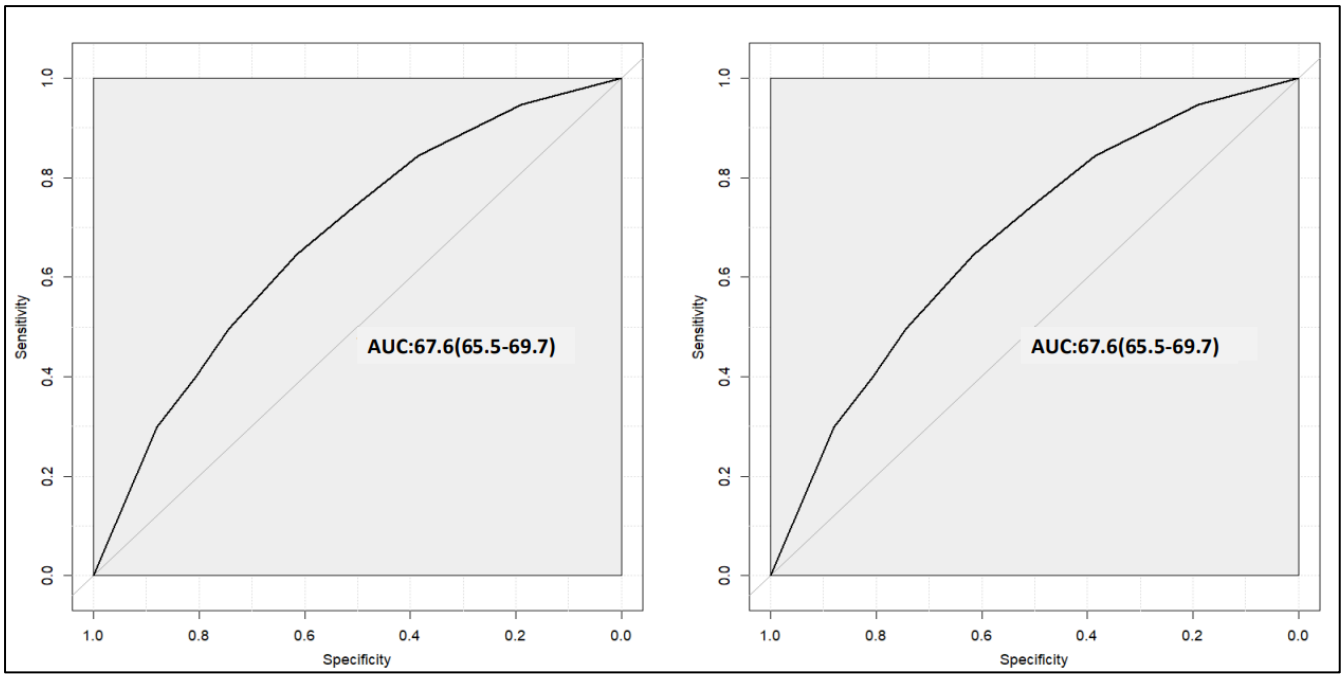
Left panel: Model using conventional risk factors. **Right panel:** Model using conventional risk factors and stroke genetic liability.

Supplementary Figure 6: ROC plot of Gradient boosting models



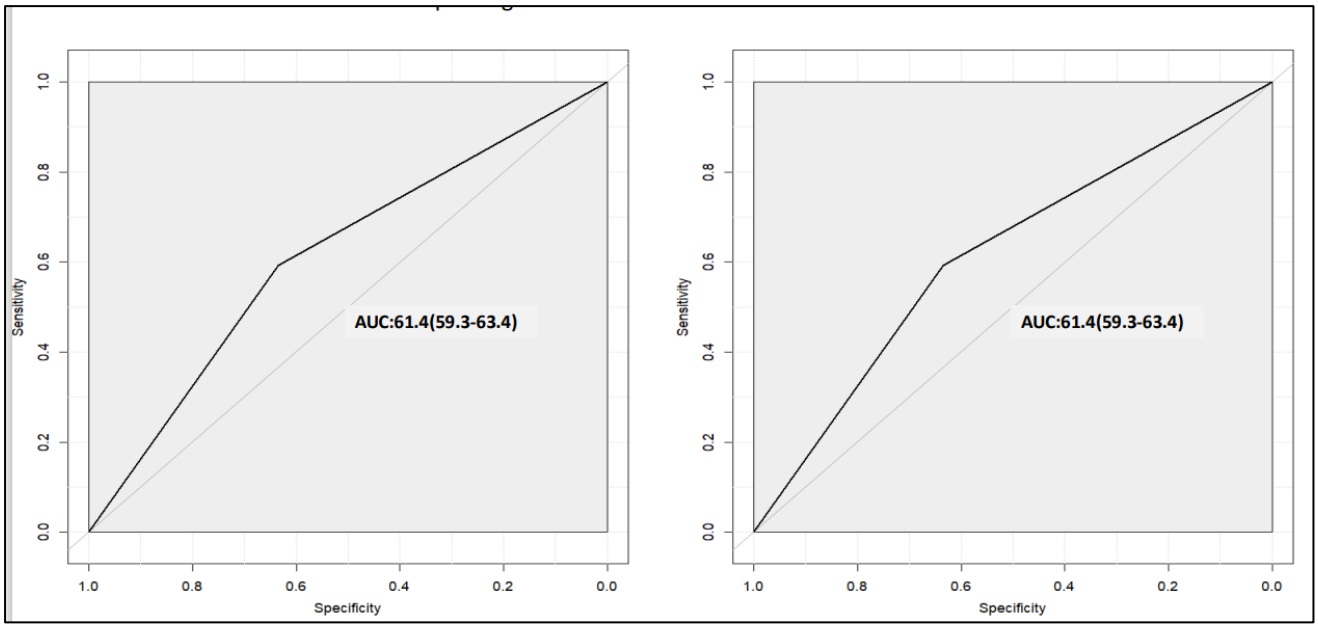
Left panel: Model using conventional risk factors. **Right panel:** Model using conventional risk factors and stroke genetic liability

Supplementary Figure 7: ROC plot of decision tree models (using pruning)



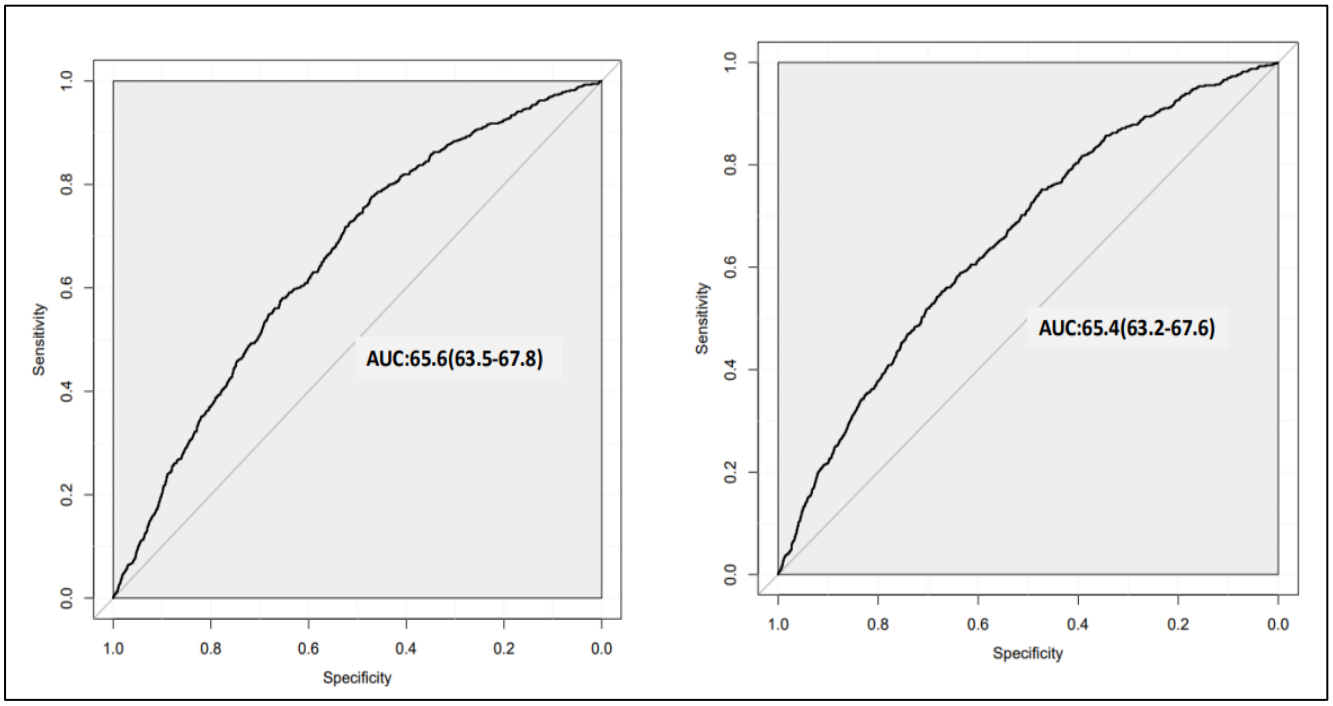
Left panel: Model using conventional risk factors. **Right panel:** Model using conventional risk factors and stroke genetic liability

Supplementary Figure 8: ROC plot of decision tree models (without pruning)



Left panel: Model using conventional risk factors. **Right panel:** Model using conventional risk factors and stroke genetic liability

Supplementary Figure 9: ROC plot of Random Forest models



Left panel: Model using conventional risk factors. **Right panel:** Model using conventional risk factors and stroke genetic liability

Paper 3

5 Chapter Five: Using discrete- and continuous-time machine learning models (Nnet, CoxNet, GLMnet) to explore sex and age differences in stroke prediction among hypertensive individuals

5.1 Introduction to Paper 3

The following chapter is based on a manuscript submitted to the Healthcare Journal for publication. The study narrowed the focus to over 100,000 hypertension participants of European ancestry from the UK Biobank. The studies examined whether incorporating stroke genetic liability could improve stroke prediction in these patients. Additionally, it examined how predictive performance differs by age and sex. The Cox proportional hazard (CoxPH) model, a penalized Cox proportional hazard model (CoxNet), a penalized logistic regression model (GLMnet), and a neural network (Nnet) were developed to estimate time-to-event risk for stroke among hypertensive patients, older and younger hypertensive patients, and hypertensive males and females separately. The area under the curve (AUC) and Brier score (BS) were used to assess the performance of the models.

healthcare

MDPI

Article

Using Discrete- and Continuous-Time Machine Learning Models (Nnet, Coxnet, Glmnet) to Explore Sex and Age Differences in Stroke Prediction Among Hypertensive Individuals

Gideon MacCarthy [1,*] and Raha Pazoki [1,2,*]

1. Department of Biosciences, College of Health, Medicine, and Life Sciences, Brunel of University London, UB8 3PH, UK, gideon.maccarthy@brunel.ac.uk (G.M.); raha.pazoki@brunel.ac.uk (R.P.)
2. Department of Epidemiology and Biostatistics, School of Public Health, St Mary's campus, Norfolk Place, Imperial College London, London W2 1PG, United Kingdom Corresponding authors: Raha Pazoki, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK, W2 1PG, raha.pazoki@brunel.ac.uk (R.P.)

* Correspondence: raha.pazoki@brunel.ac.uk

Abstract

Introduction: Stroke is one of the leading causes of death and long-term disability globally. Several studies have investigated the incidence and predictors of stroke in the healthy population; only a few have specifically focused on stroke risk prediction among individuals with hypertension. Given that hypertension is the most common modifiable risk factor for stroke, this represents an important research area to study. Improving our understanding of stroke risk among hypertensive individuals has the potential to significantly improve preventative efforts and decrease the worldwide stroke burden. **Materials and Methods:** The current study involved 116,216 participants of European ancestry from the UK Biobank. We created stroke genetic liability using data from two predictive models using clinical, lifestyle, and genetic factors (stroke genetic liability) in the training set, namely a Cox proportional hazard (CoxPH) model, a penalized Cox proportional hazard model (CoxNet), a penalized logistic regression model (GLMnet), and a neural network (Nnet), to estimate time-to-event risk for stroke among hypertensive patients, older and younger hypertensive patients, and hypertensive males and females separately. We then assessed their performances in the testing set with the area under the curve (AUC) and Brier score (BS). **Results:** The CoxPH, CoxNet, and GLMnet achieved an equal AUC of 67.0% in hypertensive patients. All models demonstrated superior performance in men compared to women, with the CoxPH model achieving the best AUC of 68% in men. All models demonstrated superior performance in older patients compared to younger patients, with the GLMnet achieving the best AUC of 61% in this group. **Conclusion:** Including stroke genetic liability in prediction models slightly improves stroke prediction for hypertensive men and older patients. The Cox proportional hazards (CoxPH) model outperformed the machine learning models in predicting stroke risk among hypertensive men.

Keywords: The Receiver Operation Characteristic (ROC), Area Under the Curve (AUC).

Brier score (BS), neural network (Nnet), Cox Proportional Hazard (CoxPH), Penalised Cox

regression (CoxNet), Penalised logistic regression (GLMnet), Hypertension, Stroke, Polygenic risk and Machine learning,

1. Introduction

Stroke is ranked among the leading causes of mortality and long-term disability in adults worldwide [1, 2]. Over 1.2 million stroke survivors live in the United Kingdom (UK). Every year, more than 100,000 people living in the UK suffer from a stroke. Between 2015 and 2035, stroke incidence is projected to increase by 60% and stroke prevalence by 120% annually [3].

Studies have shown that several risk factors, both modifiable, such as hypertension, diabetes mellitus, and chronic kidney disease, and non-modifiable, such as genetic factors, age, and sex, play a significant role in the complex mechanism of stroke occurrences [4].

According to evidence from many studies, hypertension is the most common risk factor for all types of strokes, and the majority of stroke patients have a history of hypertension [5].

Hypertension significantly increases the risk of stroke in two main ways. It damages the walls of small blood vessels in the brain, causing them to become narrower, stiffer, and more prone to plaque buildup or rupture. These damaged vessels can either lead to ischemic stroke, caused by plaque blocking blood flow to the brain, or hemorrhagic stroke, caused by the rupture of a weakened blood vessel (www.stroke.org.uk/stroke/managing-risk/high-blood-pressure, accessed on 13/08/25).

Stroke risk increases with age and varies by sex and the presence of other risk factors, such as hypertension, diabetes mellitus, atrial fibrillation, lipids, cigarette smoking, physical inactivity, chronic kidney disease, and family history [6]. Men are more likely than women to experience a stroke at a younger age. However, as women get older, especially beyond 75 years, their risk of stroke becomes higher than that of men [5, 7].

Several stroke prediction tools, including the two widely used models, the Framingham Stroke Risk Score (FSRS) and the Ischemic Cardiovascular Disease model (ICVD), have been developed with non-genetic risk factors such as age, sex, smoking, blood pressure (BP), total cholesterol, diabetes mellitus, heart disease, and body mass index (BMI) to predict a person's 10-year risk of stroke [8]. However, the two models do not integrate genetic liability into the risk estimate or include complex, nonlinear interactions among risk factors.

Recently, machine learning models have been increasingly applied to predict the risk of stroke, providing the ability to capture complex, nonlinear relationships among risk factors that traditional statistical models might not detect. Several studies [9-13] have developed machine learning models for risk of stroke prediction. These models, including the random forests, neural networks, decision trees, and gradient boosting machines, have been compared to traditional statistical approaches such as the Cox proportional hazards model. However, there is no consistency in which a prediction model is a better fit. While some studies have demonstrated superior prediction performance of machine learning models over traditional models in certain datasets, others have found only a marginal or no advantage, or the reverse. This inconsistency may be attributed to the difference in study design, study population, etc. As a result, no single machine learning or traditional model has appeared consistently superior across studies.

Although several studies have investigated the risk and predictors of stroke, only a few have specifically focused on stroke risk prediction among individuals with hypertension [14-16]. There is limited or no evidence on stroke risk prediction models for hypertensive European populations, specifically using large-scale cohorts such as the UK

Biobank. Considering that hypertension is the most common modifiable risk factor for stroke, this limitation represents an important research area to investigate. Furthermore, this research gap suggests the need for population-specific models that can improve personalized stroke risk assessment and prevention strategies in hypertensive individuals. Therefore, in this study, we aim to develop and evaluate a stroke risk prediction model for hypertensive European populations. The current study will contribute to improving the prediction of stroke risk, specifically among individuals with hypertension. Our findings may contribute to the efforts to make more accurate stroke risk classification, thereby informing clinical decision-making and public health policies driven by the reduction or prevention of hypertension-related stroke

The current study has four main objectives, including (1) assessing the predictive value of stroke genetic liability in the prediction of stroke in hypertensive individuals, (2) assessing the predictive value of stroke genetic liability in the prediction of stroke among hypertensive men and women separately, (3) assessing the predictive value of stroke genetic liability in the prediction of stroke among older and younger hypertensive patients, and (4) comparing the performance of the Cox proportional hazard model and machine learning models within the full continuous follow-up time versus discrete follow-up time.

2. Methods and Materials

2.1 Ethical Approval

The UK Biobank (UKB) received ethical approval from the Northwest Multi-Centre Research Ethics Committee as a Research Tissue Bank, and all participants provided written informed consent at the time of recruitment. The current study was conducted using UK Biobank data under approved application number 60549. Additionally, ethical approval to use secondary data from the UK Biobank for the current study was obtained from the College of Health, Medicine, and Life Sciences Research Ethics Committee at Brunel University of London (reference: 27684-LR-Jan/2021-29901-1).

2.2. Source of Study Population

Participants in this study were of European ancestry and recruited from the UK Biobank, a large, prospective cohort study. The design, recruitment procedures, and data collection methodologies have been described in detail in our previous publications [12, 17] and by Sudlow [18]. In brief, the UK Biobank recruited over 500,000 individuals aged 40–69 years between 2006 and 2010 from 22 centres across the UK. Baseline data included sociodemographic factors, lifestyle, medical history, physical measurements, and linked health records from Hospital Episode Statistics (HES) and national registries. The current study is based on hypertensive participants in the UK Biobank.

2.3. Inclusion Criteria

The current study included participants who had complete data on all relevant confounding variables. Individuals were classified as having hypertension if they met any of the following criteria: (i) a self-reported physician diagnosis of hypertension, (ii) a measured systolic blood pressure (SBP) ≥ 140 mmHg or diastolic blood pressure (DBP) ≥ 90 mmHg at baseline, or (iii) a record of using blood pressure-lowering medication either before or at the time of recruitment

2.4. Exclusion Criteria

Participants were excluded from the study based on the following criteria (**Figure 1**):

- Pregnant women and those uncertain of their pregnancy status (N = 278).

- Individuals with mismatches between self-reported and genetically determined sex (N = 320).
- Participants related to the second degree were identified using a kinship coefficient cutoff of 0.0884 (N = 33,369).
- Individuals with a prior or current diagnosis of vascular or heart problems reported at baseline (N = 25,340).
- Participants using cholesterol-lowering medications (N = 34,243).
- Individuals who reported ceasing smoking or alcohol consumption due to health reasons or medical advice (N = 58,752).
- Participants with missing data on key confounding variables (N = 61,961).
- Non-hypertensive participants were defined as those not meeting the hypertension criteria outlined in the inclusion section.

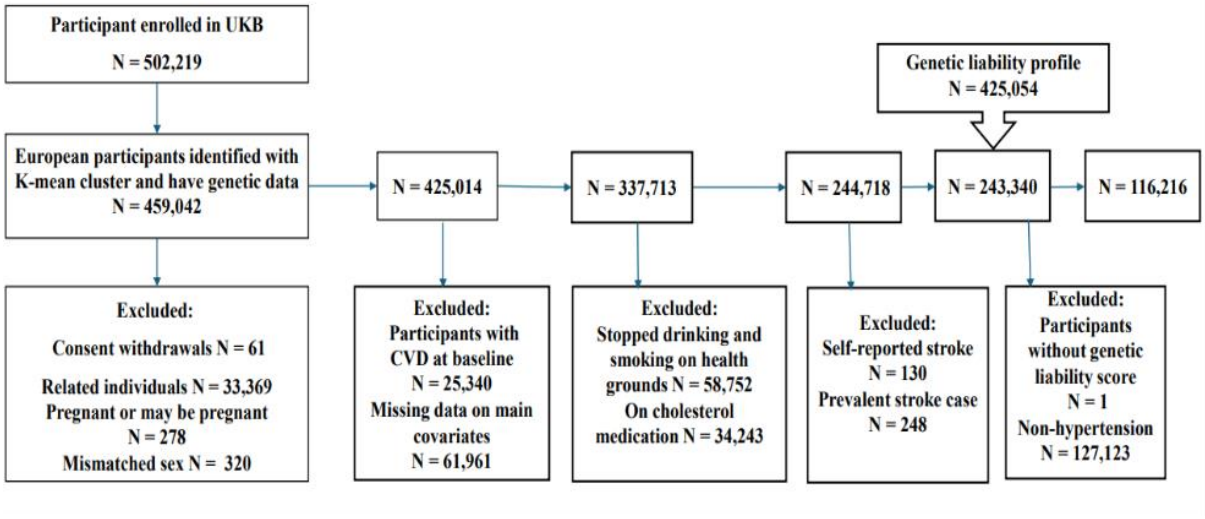


Figure 1. Exclusion Criteria of the Study: The flowchart for selecting research participants. At the start of this study, the UK Biobank (UKB) had over 500,000 participants. We employed the K-means cluster approach to extract 459,042 European-ancestry subjects with genomic data. The final dataset had 116,216 people who had hypertension and satisfied the inclusion criteria.

After applying these criteria, a final analytic sample was retained for downstream analysis.

2.5. Study Population and Period

This study focused on a subset of unrelated UK Biobank participants of European ancestry (N = 116,216; **Figure 1**) who met criteria for hypertension at baseline. The follow-up period for this study extended from the date of each participant's baseline health assessment (conducted between 2006 and 2010) to the end of March 2017. Participants who did not experience a stroke event by the end of the follow-up period were censored at that time.

2.6. Genotyping and Imputation

Details of DNA extraction, genotyping, and imputation procedures have been described in our previous manuscripts [12, 17] and [19-21]. Briefly, genotyping was performed by the UK Biobank using the UK Biobank Axiom array. Genotype imputation was conducted using the IMPUTE4 software with reference to the UK Biobank's centrally estimated genetic principal components and kinship coefficients, in accordance with the

Haplotype Reference Consortium (HRC), UK10K, and 1000 Genomes Phase 3 panels. Genetic principal components and kinship coefficients were centrally estimated by the UK Biobank to account for population stratification and relatedness [19, 21].

2.7. Study Variables

The dependent variable in this study was the incidence of stroke. The independent variables included both conventional risk factors and genetic liability to stroke (**Supplementary Table 1**). The independent variables included the following:

- Sociodemographic factors - age (in years) and sex (male/female).
- Lifestyle factors - alcohol consumption (current, previous, never) and cigarette smoking status (current, previous, never).
- Clinical factors - Body mass index (BMI), total cholesterol (TC) and low-density lipoprotein (LDL) levels, and diabetes mellitus (DM).

DM was defined based on any of the following criteria at baseline: (i) self-reported physician diagnosis of diabetes, (ii) use of insulin or other glucose-lowering medication, (iii) hemoglobin A1c (HbA1c) ≥ 48 mmol/mol (6.5%), or (iv) fasting glucose level ≥ 7.0 mmol/L [22].

2.8. Definition of the Outcome

The primary outcome was the incidence of stroke, as defined by the International Classification of Diseases, 10th Revision (ICD-10 codes I60–I67). Stroke events were identified using Hospital Episode Statistics (HES) records and death registries, capturing first-ever stroke events with the following ICD-10 codes: subarachnoid haemorrhage (I60.0–I60.9), intracerebral haemorrhage (I61.0–I61.9), cerebral infarction (I63.0–I63.9), stroke not specified as haemorrhage or infarction (I64), occlusion and stenosis of pre-cerebral and cerebral arteries (I65.0–I65.9, I66.0–I66.9), and other cerebrovascular diseases (I67.0–I67.9). The follow-up period was calculated from each participant's baseline health assessment date to the first stroke event or the end of follow-up (March 31, 2017). Participants who did not experience a stroke during follow-up were censored at that point.

2.9. Computation of Genetic Liabilities

The selection of single-nucleotide polymorphisms (SNPs) and the preprocessing steps have been described in genome-wide association studies, with the primary source being [23], which focused on stroke genetics in European ancestry populations (S**upplementary Data S1**). Genetic liability to stroke was quantified using a polygenic risk score (PRS), computed with PLINK 1.9. The PRS was calculated as a weighted sum of stroke-associated risk alleles, where the weights corresponded to the SNP effect sizes (β coefficients) reported in the reference GWAS. Key quality control procedures included:

- Filtering SNPs with low call rates or violations of the Hardy-Weinberg equilibrium,
- Excluding SNPs with minor allele frequency (MAF) < 1%,
- Linkage disequilibrium (LD) pruning using a window size of 250 kb, step size = 50, and an r^2 threshold of 0.1.

After LD pruning, 252,903 SNPs were retained for PRS calculation (**Supplementary Figure 1**). The PRS was computed using the *"score"* function in PLINK. The resulting PRS and all quantitative independent variables were standardized using the "scale" function in the R package and included as independent variables in regression models to assess the association between genetic predisposition to stroke and incident stroke among individuals with hypertension.

2.10. Statistical Analyses

The statistical analysis approach used in this study follows the methodology described in our previous manuscript [12]. Briefly, baseline characteristics were summarized using descriptive statistics, including frequencies and percentages for categorical variables and means with standard deviations for continuous variables. Differences between stroke and non-stroke groups were assessed using chi-square tests for categorical variables and Wilcoxon rank-sum tests for continuous variables. Baseline characteristics of the study population were summarized using the *gtsummary* and *table1* packages in the R program.

To assess the relationship between the predictors and incident stroke, we used univariable and multivariable Cox proportional hazards regression models (CoxPH). Before multivariable modeling, we checked for multicollinearity among quantitative variables by using Pearson correlation coefficients with the *"cor"* function. Highly correlated variables ($r^2 \geq 0.80$) were identified with the "*findCorrelation"* function from the caret package in R and visualized using the *ggcorrplot* package. Variables with high collinearity and limited univariable association with the outcome were excluded to prevent overfitting.

To examine differences in stroke-free survival time among hypertensive participants with varying levels of genetic liability, we performed Kaplan–Meier survival analysis over the defined follow-up period. The survival functions were stratified by categories of genetic liability to stroke (low, intermediate, and high) and compared using the log-rank test (**Supplementary Figure 2**).

To study the effect of the stroke genetic liability while adjusting for the impact of different covariates, sequences of multivariate CoxPH models were created.

Model 1 included only stroke genetic liability; model 2 adjusted for age and sex; model 3 included adjustments for age, sex, diabetes mellitus, and total cholesterol levels; and model 4 included adjustments for age, sex, diabetes mellitus, and total cholesterol, as well as smoking and alcohol status.

The results of the multiple Cox proportional hazards regression analysis are presented as adjusted hazard ratios (HRs) along with the corresponding 95% confidence intervals (CIs). A significance level of $p < 0.05$ indicates that a predictor is independently associated with the incidence of stroke. A forest plot was employed to visualize the hazard ratios for all covariates using the *"forest_model"* function from the *forestmodel* package.

The Cox proportional hazards assumption was evaluated using the global Schoenfeld residuals test, and individual variable assessments were visualized. Meeting the proportionality assumption indicates the validity of the Cox regression results.

All analyses were performed using R software (version 4.4.1) for Windows (R Development Core Team, 2010).

2.11. Prediction Model Development and Performance Assessment

In this study, sets of integrated prediction models were developed to predict the incidence of stroke. The integrated prediction model combines genetic liability for stroke (genetic risk) and the conventional risk factors (**Supplementary Figure 3**)

Using the "*createDataPartition*" function from the *caret* package, we randomly partitioned our dataset into a training set (70%; N = 81,352; Event = 959 and Non-event = 80,393) and a testing set (30%; N = 34,864; Event = 417 and Non-event = 34,447). The training data was used to create predictive models, while the testing data was utilized to assess the models' performance.

2.12. Implementation

To assess the added predictive value of stroke genetic liability (SGL) for predicting stroke in hypertensive individuals, we implemented and compared various statistical and machine learning models using UK Biobank data. We implemented both the continuous-

and discrete-time survival models to predict the risk of stroke incidence in hypertensive individuals. The continuous-time models included standard Cox proportional hazards (CoxPH) and penalized Cox regression (Coxnet). For penalized logistic regression (GLM-Net) and a neural network-based discrete survival model (Nnet), we adopted the discrete-time survival approach. The packages, functions, and parameters used in the implementation of the models are presented in **Supplementary Table 2**. The core concept behind prediction in the discrete-time survival framework is to develop models that estimate the probability of survival within each discrete time interval. By treating the occurrence of an event in each interval as a binary outcome, the task can be framed as a sequence of binary classification problems [24].

In this study, we implemented discrete-time survival analysis by discretizing the continuous follow-up time into defined intervals. This transformation allowed us to restructure the data into a person-period (long) format, in which each row represents a unique individual-time interval combination. This approach ensures that any change in an individual's status during the follow-up period could be accurately captured. The transformation was conducted using the *"discSurv"* function from the *discSurv* package in R. After reshaping the dataset, we addressed class imbalance, an inherent issue in survival data due to the relatively low incidence of stroke events. To mitigate this, we employed the ROSE (Random Over-Sampling Examples) method using the ROSE function from the ROSE package in R. The use and justification for ROSE were detailed in our previous manuscript [17]. In brief, ROSE generates synthetic examples of the minority class and under-samples the majority class to create a more balanced training set [25]. This pre-processing ensured that the discrete-time models were trained on a balanced dataset, thereby improving model robustness and predictive power, particularly for rare outcomes such as stroke [26].

Cox Proportional Hazards Model (CoxPH)

Cox proportional hazards regression [27, 28] is a popular statistical approach for analysing survival data and determining the association between the time until an event (such as death, failure, or illness recurrence) occurs and one or more predictors. We implemented the Cox proportional hazard models using the *"coxph"* function from the *Survival* package in R software.

In addition to the previously described models, we implemented the following machine learning models in a continuous-time framework: a penalized Cox regression model (CoxNet) and two discrete-time survival models, a penalized logistic regression model, and a neural network, to further explore stroke risk prediction.

Penalized Regression Models

In comparison to traditional (unpenalized) regression methods, such as Cox and logistic regression, penalized regression models improve prediction performance on new data by applying regularization. Regularization introduces a penalty by shrinking the size of less informative coefficients towards zero and only retaining those with coefficients greater than zero [29, 30]. The penalized method employed in this study was elastic-net, which is a combination of LASSO (L1) and RIDGE (L2) regularization.

1. Penalized Cox Regression (CoxNet)

We utilized an elastic-net regularized Cox regression model with the *"cv. glmnet"* function in the *glmnet* package. The model was trained using 10-fold cross-validation and evaluated by the area under the receiver operating characteristic curve (AUC). The optimized CoxNet model was generated by tuning parameters alpha = 0.5 and lambda = 0.0001 **(Supplementary Table 2).**

2. Penalized Logistic Regression (GLMnet)

We applied an elastic-net regularized logistic regression model using the *"cv. glmnet"* function in the *glmnet* package. The model was trained using 10-fold cross-validation and evaluated by the area under the receiver operating characteristic curve (AUC). The optimized GLMnet model was generated by tuning parameters alpha = 0.5 and lambda = 0.002 **(Supplementary Table 2).**

Neural Network Model (Nnet)

A feedforward neural network was trained using the Nnet method within the caret framework. The neural net was described and implemented in our previous manuscript [17]. The survival time was discretized into intervals, and the model was evaluated using AUC under 10-fold cross-validation. The optimized neural network model was generated by tuning parameters size = 5 and decay = 0.1 (**Supplementary Table 2).**

The predictions from all the models were obtained using the "*predict"* function. In this study, we additionally focused on probability calibration to improve the interpretability and reliability of stroke risk predictions. To calibrate the predicted probabilities, we applied Platt scaling, also known as the sigmoid method [31],[31] to transform model-generated scores into calibrated probability estimates. This post-processing technique involves fitting a logistic regression model where the model-predicted probabilities are used as the independent variable, and the binary outcome (stroke occurrence) is the dependent variable [32]. The resulting transformation produces rescaled probability estimates that more accurately reflect the true event rates, thereby enhancing clinical interpretability. Details have been described in our previous manuscripts [12, 17].

2.13. Assessing model prediction performance

We assessed the predictive performance of each model in the testing dataset using the receiver operating characteristic curve (AUC) and the Brier Score (BS). We assessed the model's discrimination ability using the AUC. Model calibration and overall performance were assessed using the BS, which measures the mean squared difference between predicted probabilities and actual outcomes; lower BS values indicate better calibration and prediction accuracy [32] of these metrics can be found in our previous work [12, 17].

For the continuous-time survival models, CoxPH and CoxNet, we used the "*rcorr.cens"* function from the *Hmisc* package to calculate the concordance index (c-index or AUC), and for the discrete-time survival models, GLMnet and Nnet, the AUC was calculated using the "*roc"* function from the *pROC* package. Furthermore, we evaluated the discriminative performance of all models specifically in hypertensive individuals at various time points (2, 4, 6, 8, and 10 years) using the "*timeROC"* function from the *timeROC* package. ROC curves were generated for each model using the *"plot"* function.

3. Results

3.1. Study Characteristics and Statistical Analysis

Table 1 presents the baseline characteristics of the study population, composed of 116,216 unrelated participants of European ancestry with hypertension, including 13,766 incident cases and 102,450 controls. Key demographic and clinical features significantly differ between the two groups.

Table 1. Baseline Characteristics of Study Population Stratified for Stroke Event and Non-Stroke Event among Hypertension Patients with UK Biobank Participants (N=116,216).

Characteristics	Overall (N=116,216)	Control (N=102,450)	Case (N=13,766)	HR (95% CI)	p-value

Age(years); Mean(SD)	57.6 (7.53)	57.6 (7.53)	60.9 (6.70)	1.67 (1.57, 1.78)	<0.001
Body mass index (kg/m^2); Mean (SD)	28.0 (4.83)	28.0 (4.83)	27.9 (4.77)	0.98 (0.93, 1.03)	0.44*
Total cholesterol (mmol/L); Mean (SD)	6.05 (1.06)	6.05 (1.06)	5.96 (1.07)	0.92 (0.88, 0.97)	0.004
Low-density lipoprotein (mmol/L); Mean (SD)	3.84 (0.81)	3.84 (0.81)	3.81 (0.81)	0.96 (0.91, 1.02)	0.16*
Sex (Male); n (%)	55269 (47.6%)	54476 (47.4%)	793 (57.6%)	1.50 (1.35, 1.67)	<0.001
Diabetes Mellitus (Yes); n (%)	4638 (4.0%)	4545 (4.0%)	93 (6.8%)	1.75 (1.42, 2.16)	<0.001
Smoking status					<0.001
Current n (%)	37098 (31.9%)	36579 (31.9%)	519 (37.7%)	1.33 (1.19, 1.48)	
Previous; n (%)	1320 (1.1%)	1283 (1.1%)	37 (2.7%)	1.15 (0.83, 1.59)	
Drinking status					<0.001
Current; n (%)	108867 (93.7%)	107635 (93.7%)	1232 (89.5%)	0.61 (0.49, 0.78)	
Previous; n (%)	3333 (2.9%)	3263 (2.8%)	70 (5.1%)	2.67 (1.92, 3.72)	
n (%); Mean (SD), HR = Hazard Ratio, CI = Confidence Interval					

The p-value is from a univariate analysis of the Cox Proportional Hazard Model comparing the distribution of the baseline characteristics among stroke events and non-stroke events in hypertension patients. *Not significant; **DM** = Diabetes Mellitus, **HR** = Hazard Ratio, CI = Confidence Interval, **SD** = Standard Deviation. .

Participants who developed a stroke during the study were, on average, older than controls (mean age 60.9 vs. 57.6 years). Age was strongly associated with incident stroke (HR = 1.67, 95% CI = [1.57, 1.78], *p-value* < 0.001). Male participants had a significantly higher hazard than female participants (HR=1.50, 95% CI = [1.35, 1.67], *p-value* < 0.001). Among lifestyle factors, current smoking was associated with a higher risk of stroke (HR = 1.33, 95% CI = [1.19, 1.48], *p-value* < 0.001), while current alcohol consumption was inversely associated with the risk of stroke (HR = 0.61, 95% CI = [0.49, 0.78], *p-value* < 0.001).

Among the clinical factors, diabetes mellitus was also strongly associated with an increased risk of incident stroke (HR = 1.75, 95% CI = [1.42, 2.16], *p-value* < 0.001). Total cholesterol was marginally protective (HR = 0.92, 95% CI: 0.88–0.97, p = 0.004), whereas low-density lipoprotein (LDL) cholesterol and BMI were not significantly associated with stroke risk (*p-value* = 0.16 and 0.44, respectively).

The correlation matrix plot showed that total cholesterol (TC) and low-density lipoprotein (LDL) are highly correlated, and the stroke genetic liability is independent of the other covariates (**Figure 2**)

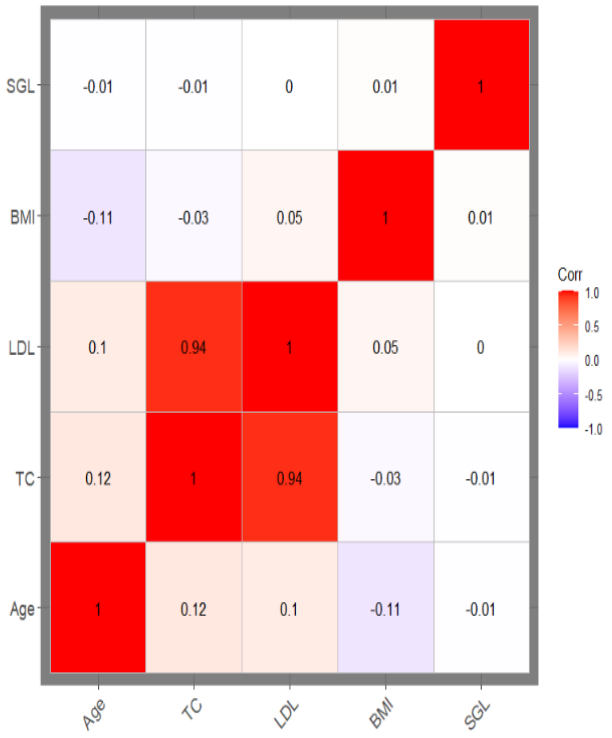


Figure 2. Correlation matrix plot: The plot shows the correlation coefficients between numerical features. LDL and TC are highly correlated ($r^2 > 0.8$). LDL was excluded from further analysis (prediction model construction). BMI: Body mass index; TC: Total cholesterol; LDL: Low-density lipoprotein cholesterol; SGL: Stroke genetic liability.

Table 2 shows the results of a sequence of Cox proportional hazards models evaluating the relationship between stroke genetic liability (as a continuous variable) and incident stroke, from univariable (Model 1) to adjusted multivariable models (Models 2–4). Genetic liability was consistently associated with a statistically significant increased risk of stroke.

Table 2. Univariable and Multivariable Cox Proportional Hazard Analysis Assessing the Association of Stroke Genetic Liability with Stroke.

Characteristic	HR	95% CI	p-value
Continuous Stroke Genetic Liability			
Model 1 (unadjusted)	1.11	1.06, 1.18	<0.001
Model 2 (Model 1 + age and sex)	1.12	1.06, 1.18	<0.001
Model 3 (Model 2 + clinical factors) *	1.12	1.06, 1.18	<0.001
Model 4 (Model + lifestyle factors)**	1.12	1.06, 1.18	<0.001

HR = Hazard Ratio; CI = Confidence Interval. Model 1: univariable Cox proportional hazard analysis of stroke genetic liability. Model 2: Adjusted for Age and Sex. *Model 3 included adjustments for age, sex, diabetes mellitus, and total cholesterol levels. **Model 4 included adjustments for age, sex, diabetes mellitus, and total cholesterol levels.

In the unadjusted model (Model 1), the hazard ratio (HR) was 1.11 (95% CI = [1.06, 1.18], *p-value* < 0.001), indicating that for each unit (standard deviation) increase in genetic liability, the risk of stroke increased by approximately 11%. This association remained steady after adjustment for age and sex in Model 2 (HR = 1.12, 95% CI = [1.06, 1.18], *p-value* < 0.001) and was unchanged with further inclusion of conventional risk factors such as diabetes, body mass index, and total cholesterol in Model 3. Even after including

additional covariates such as smoking status and alcohol status in Model 4, the effect estimate remained stable (HR = 1.12, 95% CI = [1.06, 1.18], *p-value* < 0.001).

These results suggest that stroke genetic liability is an independent risk factor for stroke risk, showing a consistent and statistically significant effect across all levels of adjustment in the models. The assessment of the Cox proportional hazards assumption indicated that all the predictor variables individually achieved a *p-value* > 0.05 based on Schoenfeld residuals, and the global test also established that the overall model satisfied the assumption (*p-value* > 0.05). These results suggest that the Cox proportional hazards assumption was not violated for any predictor variable or the model as a whole (**Supplementary Table 3 and Supplementary Figure 4**).

3.2. Prediction Performance of the Models

Table 3 presents the results of performance from a series of prediction model techniques employed in this study

Table 3. The Results of the Model Performance for Incident Stroke Prediction in the Hypertensive Patients in the UK Biobank.

Models	Model Features	Sample	AUC (95% CI)	Brier Score
CoxPH				
	Conventional factors + genetics	whole population (n =116,216)	67.0 (64, 70)	0.01
	Conventional factors + genetics	Men (n = 55,269)	68.0 (64, 71)	0.01
	Conventional factors + genetics	Women (n = 60,947)	65.0 (60, 69)	0.01
	Conventional factors + genetics	Age > 59 (n = 55370)	60.0 (56, 64)	0.02
	Conventional factors + genetics	Age < 59 (n = 60846)	58.0 (53, 63)	0.02
CoxNet				
	Conventional factors + genetics	whole population (n=116,216)	67.0 (64, 70)	0.01
	Conventional factors + genetics	Men (n = 55,269)	67.0 (64, 71)	0.01
	Conventional factors + genetics	Women (n = 60,947)	65.0 (60, 69)	0.01
	Conventional factors + genetics	Age > 59 (n = 55370)	60.0 (57, 64)	0.02
	Conventional factors + genetics	Age < 59 (n = 60846)	56.0 (51, 61)	0.01
GLMnet				
	Conventional factors + genetics	whole population (n=116,216)	67.0 (64, 70)	0.001
	Conventional factors + genetics	Men (n = 55,269)	67.0 (63, 70)	0.002
	Conventional factors + genetics	Women (n = 60,947)	65.0 (61, 69)	0.001
	Conventional factors + genetics	Age > 59 (n = 55370)	61.0 (58, 65)	0.002
	Conventional factors + genetics	Age < 59 (n = 60846)	57.0 (52, 62)	0.001

Nnet				
	Conventional factors + genetics	whole population (n=116,216)	66.0 (63, 69)	0.001
	Conventional factors + genetics	Men (n = 55,269)	65.0 (62, 69)	0.002
	Conventional factors + genetics	Women (n = 60,947)	64.0 (59, 68)	0.001
	Conventional factors + genetics	Age > 59 (n = 55370)	60.0 (56, 63)	0.002
	Conventional factors + genetics	Age < 59 (n = 60846)	55.0 (50, 60)	0.001

Across the full study sample (N = 116,216), the inclusion of genetic liability of stroke yielded small but consistent improvements in the risk of stroke prediction.

Both the Cox proportional hazards (CoxPH) and penalized Cox regression (CoxNet) achieved an equal AUC of 67.0 (95% CI = [64, 70], **Supplementary Figure 5**). The Brier score (BS) for both models was 0.01. The penalized logistic model (GLMnet) also achieved an AUC of 67.0 (95% CI = [64, 70], **Supplementary Figure 6**), while the neural network-based model (Nnet) achieved an AUC of 66.0 (95% CI = [63, 69**], Supplementary Figure 6**). Both GLMnet and Nnet achieved a BS of 0.001. The AUC and BS values indicate good prediction performance across all the models.

Our analysis of hypertensive men and women separately shows that the models achieved higher discrimination ability (AUC) in hypertensive men than in hypertensive women.

Among hypertensive men (n = 55,269), the CoxPH achieved an AUC of 68.0 (95% CI = [64, 71]) and a BS of 0.01. The model achieved an AUC of 67.0 (95% CI = [64, 71]) and BS of 0.01 for CoxNet, an AUC of 67.0 (95% CI = [63, 70]) and BS of 0.002 for GLMnet. The Nnet attained an AUC of 65.0 (95% CI = [62, 69]) and a BS of 0.002. Among hypertensive women participants (n = 60,947), both the CoxPH and CoxNet achieved AUCs of 65.0 (95% CI = [60, 69]) and BS values of 0.01. GLMnet achieved an AUC of 65.0 (95% CI = [61, 69]) and BS of 0.001, while Nnet achieved an AUC of 64.0 (95% CI = [59, 68]) and BS of 0.001

The models also showed better discrimination for participants who were older than the study sample's average age (59 years) than for younger participants. In older participants (n=55,370), the GLMnet achieved an AUC of 61.0 (95% CI = [58, 65]) and BS of 0.002, while CoxPH, CoxNet, and Nnet achieved a similar AUC of 60.0.

3.3. Model Performance at Multiple Follow-up Time Points

Table 4 presents the discriminative abilities of the models at different follow-up time points (2, 4, 6, 8, and 10 years) for the whole study sample. The result shows that all the continuous-time survival models' performances improved with time. CoxPH and CoxNet performed similarly and consistently (approx. 65 to 67 AUC) across the follow-up time points. CoxNet slightly improved with time and performed best in year 8 (AUC = 67.37, **Supplementary Figure 7**). The discrete-time survival models (GLMnet and Nnet) performed strongly in years 4 and 6, then degraded after. Both GLMnet and Nnet achieved their highest AUCs in year 4 (AUC = 69.98 and 68.78, respectively; **, Supplementary Figure 8**), suggesting strong mid-term prediction.

Table 4. Model Performance at Multiple Follow-Up Time Points in the Testing Set.

Years	Cases	Survivors (without the event)	Censored	CoxPH (AUC %)	Coxnet (AUC %)	GLMnet (AUC %)	Nnet (AUC %)
2	75	34789	0	64.62	64.63	63.18	62.47
4	168	34696	0	65.26	65.29	69.98	68.78
6	282	34582	0	65.91	65.91	68.68	67.10
8	389	20450	14025	67.35	67.37	66.72	65.21
10	417	0	34447	N/A	N/A	50.72	53.40

Table 4 shows the time-dependent ROC AUC (Area Under the Curve) estimates over different years (2, 4, 6, 8, and 10) using IPCW (Inverse Probability of Censoring Weighting) for six predictive models: CoxPH, CoxNet, GLMnet, and Nnet.

4. Discussions

This large-scale cohort study investigated the predictive value of genome-wide stroke genetic liability derived from 252,903 stroke-associated SNPs in over 116,000 hypertensive individuals of European ancestry. In the current study, we constructed machine learning models and assessed their ability to predict stroke incidents in all hypertensive participants, stratified by age and sex, using a survival analysis framework. Our models included traditional risk factors of stroke alongside genetic liability. To examine the discriminatory accuracy of time-to-event data models, we employed both discrete- and continuous-time machine learning algorithms. These included Elastic Net regression models, specifically the penalized Cox model (CoxNet) and penalized logistic regression (GLMnet), as well as a neural network (Nnet). These models were compared to the Cox proportional hazards (CoxPH) model, which is used as the gold standard for survival analysis.

Our main findings were (1) in all hypertension patients, all the models (both discrete and continuous-time survival models) showed modest discriminating value, with AUC ranging from 66 to 68. CoxNet and GLMnet achieved equal discrimination ability as the CoxPH model in the follow-up period. The discrete-time survival models (GLMnet and Nnet) outperformed the continuous-time models at narrower follow-up intervals (2-4 years and 4-6 years from baseline). (2) All models consistently achieved higher AUC values in hypertensive men than in hypertensive women. (3) All models consistently achieved higher AUC values in older participants (over the median age of 59 years old) than in younger participants.

In comparison with our previous studies on stroke risk prediction [12], the current study (1) focuses on hypertensive individuals. (2) While our previous work focused on models using continuous-time survival models, the current work adds discrete-time machine learning models (GLMnet, a form of penalized logistic regression model, and Nnet). (3) Discretized the follow-up time and incorporated it as a covariate within the discrete time machine learning models, and (4) Our present work additionally extended and leveraged the readily available open-source software packages in the R program for the binary classification problem to survival analysis frameworks. That is, we reformulated the continuous-time survival model as a binary classification problem. This approach overcomes several limitations of traditional continuous-time survival models, including problems with tied event times, the proportional hazard assumption requirement. Discrete-time survival models offer practical alternatives to continuous-time models, particularly when the exact event times are unknown and only the time interval during which the

event occurs is known. This framework restructures survival analysis as a series of binary classification problems, enabling a wide range of classification algorithms to be used for estimating conditional survival probabilities [33].

While machine learning methods are appropriate for capturing complex, non-linear relationships between predictors and outcomes, our findings suggest that the Elastic Net models (GLMnet and CoxNet) performed similarly to our reference regression model (CoxPH) for stroke prediction in hypertensive patients. In contrast, the more complex model, Neural Network (Nnet), demonstrated inferior performance relative to CoxPH. This observation suggests that the CoxPH model may be sufficient for developing an effective and accurate stroke risk prediction model in hypertensive individuals from the UK Biobank (UKB). This observation could be due to the absence of non-linear or complex interactions among the predictor variables used in the current study. This finding is consistent with the results of our previous work [12], where we demonstrated that CoxPH outperformed all the machine learning models, including Random Forest (RF), the Gradient Boosting Model (GBM), and the Decision Tree (DT). However, in our previous work, we did not consider the binary classification machine learning method and did not explore age and sex strata.

4.1. Comparison between traditional and machine learning models that include the effect of time.

The performance of machine learning techniques has been compared to traditional Cox regression models in predicting stroke risk; however, the results in the literature are inconsistent. For example, two previous studies, Chen, Y [11] and Chun [9], each created ML-based prediction models and compared their performance against Cox regression. While both studies highlighted the potential of ML techniques, their conclusions about superiority to Cox regression were not conclusive.

Chun and colleagues [9] developed a prediction using age, hypertension, coronary heart disease, diabetes, and smoking and observed that for the 9-year risk of stroke prediction in Chinese adults. They observed that the Cox regression model and the machine learning models had similar prediction performance. The Cox regression model marginally outperformed the Random Survival Forest model but performed worse than the Gradient Boosted Trees.

For predicting 30-day stroke readmission in a Taiwanese cohort, Chen, Y, and colleagues included a wide range of clinical, demographic, and pre-rehabilitation functional status scores in their prediction models. They observed that among the machine learning models, only the neural network and the random forest models outperformed the Cox regression model. These results differed from the findings of our current study, which found that both the neural network and random forest models underperformed in comparison to the CoxPH model.

Unlike our study, the two previous studies were conducted in Asian populations, whose genetic architectures differ from the European population we studied. While Chen [11] explored the prediction of 30-day stroke readmission and Chun [9] focused on the prediction of first-ever incident stroke within a population-based cohort, our study focused on predicting first-time stroke events in hypertensive Europeans beyond the 30-day window and uniquely incorporated stroke genetic liability as a predictor. The methodological divergences highlight the uniqueness of our approach and may explain differences in prediction accuracy and interpretability between studies. These fundamental differences in research methodology, outcome definitions, and study population restrict direct comparisons of our findings with previous studies and may explain the difference in model performances.

4.2. Comparison based on the different follow-up time lengths

When we evaluated model performance across different follow-up intervals for hypertensive individuals, we observed that during shorter follow-up intervals (2–4 years and 4–6 years from baseline), all the discrete-time survival models outperformed the continuous-time survival models. Within shorter follow-up intervals, the GLMnet was the best-performing model, followed by the Nnet model. In contrast, with a longer follow-up interval (6–8 years), the continuous-time survival models performed better, with Cox proportional hazards (CoxPH) and CoxNet emerging as the best-performing models. These observations were consistent with the findings of Chun [9], who demonstrated that the binary classification machine learning models, including logistic regression (LR), support vector machines (SVM), gradient boosting trees (GBT), and multilayer perceptron (MLP), outperformed survival models such as Cox regression and random survival forests. They reported that the best performances in predicting stroke were observed within shorter follow-up intervals (0–3 years, 3–6 years, and 6–9 years from baseline). This implies that for prediction within shorter follow-up times, discrete-time survival models (GLMnet and Nnet) are more suitable for risk prediction.

4.3. Gender differences

Across all models, both machine learning and traditional models, the model's discrimination ability was consistently better among hypertensive men compared to hypertensive women. Among hypertensive men, the CoxPH model provided better discriminative accuracy than machine learning models. These findings diverged from the previous study by Chun [9].

Within a population-based Chinese cohort, Chun [9] evaluated the prediction performance of machine learning algorithms with the Cox regression model for predicting the first-ever stroke in men and women separately. They observed that the models improved stroke prediction in women more than in men, with GBT providing the best discrimination (AUC: 0.833 in men, 0.836 in women) and calibration. This observation deviates from the findings of our current study, in which both machine learning and the Cox model improved stroke prediction in hypertensive men more than in hypertensive women, with the CoxPH providing the best discrimination (AUC: 68% in men, 65% in women). This further highlights the possible impact of population characteristics, outcomes, and study methodology on predictive model performance.

4.4. Age differences

Across all models, both machine learning and traditional models, the discrimination performance of the models consistently achieved a better discrimination value among older hypertensive participants compared to younger hypertensive participants. Among older hypertensive participants, the GLMnet model provided better discriminative accuracy. Our finding was consistent with findings from several studies, including Ding [34] and Gong [14], who demonstrated that the risk of stroke increases with age, especially among older hypertensive patients. These observations show the potential age-based differences in model performance and/or underlying risk factor patterns.

The general clinical implication of our study is that the absolute risk estimates derived from the stroke genetic liability-enhanced model could help identify older hypertensive patients, especially hypertensive men, at high risk of stroke, who can then be prioritized for preventive interventions, including the start of pharmacological treatments. A study of 100,000 UK adults [35] found that the small improvement in C-index by polygenic risk scores for cardiovascular disease (CVD) could translate to a 7% increase in CVD event prevention compared to conventional risk factors alone.

4.5. Strengths and Limitations

This study has several strengths. First, this is the first study to employ both discrete- and continuous-time survival models to evaluate the added predictive value of stroke genetic liability and predict risk of stroke in hypertensive individuals of European ancestry. Second, our study utilized a large sample size and a long follow-up time to accurately identify stroke incident cases. However, there are some limitations in this study. First, we incorporated a few predictor variables into our models. Second, data on lifestyle factors such as drinking and smoking behaviours are self-reported and may be inaccurate. Third, the small number of incident stroke cases in our study may have an impact on the predictive power of the models. Therefore, we recommend that future studies should include a large number of stroke occurrence cases and a broader set of predictor variables to improve the efficacy of the machine learning models over the CoxPH model. Fourth, the analyses in this study were restricted to UK Biobank participants with European ancestry and to participants who have been identified as or diagnosed with hypertension. These factors might limit the generalisation of our findings to a wider UK population from different ancestries or health statuses.

5. Conclusion

In conclusion, within the survival analysis framework, the Cox proportional hazards (CoxPH) model demonstrated superior performance in predicting stroke risk among hypertensive patients compared to the machine learning models evaluated. Our findings may provide stroke risk stratification and prevention of stroke in the hypertensive population.

The inclusion of stroke genetic liability in the models improved stroke prediction for a small percentage of the population. Therefore, its application in clinical practice is uncertain. Conventional risk factors may still have more influence on the prediction of stroke in hypertensive patients. The findings suggest that genetic liability alone has limited predictive value for most people, but they might still have a role in highly targeted interventions. In terms of cost effectiveness, given that only a small percentage of the population benefits from genetic risk scores, in future health economics studies are needed to establish if the costs of genetic testing could outweigh the potential improvements in stroke prevention in hypertensive patients.

Supplementary Materials: Supplementary Data S1- Supplementary Data S1: List of genetic variants summary statistics used to construct the genetic risk scores.

Author Contributions: Conceptualization, R.P.; Data curation, Formal analysis, G.M.; Investigation, G.M.; Methodology, G.M., R.P.; Project administration, G.M., and R.P.; Resources, R.P.; Supervision, R.P; Writing— original draft, G.M.; Writing—review & editing, G.M., and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: R.P. and GM were supported by the Brunel University of London BRIEF award. The MEGASTROKE project received funding from sources specified at http://www.megastroke.org/acknowledgments.html

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Brunel University London, College of Health, Medicine, and Life Sciences (27675-LR-Feb/2021- 31174-2, 5 February 2021).

Informed Consent Statement: Patient consent was waived due to the current study was conducted using UK Biobank data under approved application number 60549.

Data Availability Statement: Not applicable.

Acknowledgments: This research has been conducted using the UK Biobank Resource under Application Number 60549.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G, Ahmed M, Aksut B, Alam T, Alam K, Alla F, Alvis-Guzman N, Amrock S, Ansari H, Ärnlöv J, Asayesh H, Atey TM, Avila-Burgos L, Awasthi A, Banerjee A, Barac A, Bärnighausen T, Barregard L, Bedi N, Belay Ketema E, Bennett D, Berhe G, Bhutta Z, Bitew S, Carapetis J, Carrero JJ, Malta DC, Castañeda-Orjuela CA, Castillo-Rivas J, Catalá-López F, Choi JY, Christensen H, Cirillo M, Cooper L, Criqui M, Cundiff D, Damasceno A, Dandona L, Dandona R, Davletov K, Dharmaratne S, Dorairaj P, Dubey M, Ehrenkranz R, El Sayed Zaki M, Faraon E, Esteghamati A, Farid T, Farvid M, Feigin V, Ding EL, Fowkes G, Gebrehiwot T, Gillum R, Gold A, Gona P, Gupta R, Habtewold TD, Hafezi-Nejad N, Hailu T, Hailu GB, Hankey G, Hassen HY, Abate KH, Havmoeller R, Hay SI, Horino M, Hotez PJ, Jacobsen K, James S, Javanbakht M, Jeemon P, John D, Jonas J, Kalkonde Y, Karimkhani C, Kasaeian A, Khader Y, Khan A, Khang YH, Khera S, Khoja AT, Khubchandani J, Kim D, Kolte D, Kosen S, Krohn KJ, Kumar GA, Kwan GF, Lal DK, Larsson A, Linn S, Lopez A, Lotufo PA, El Razek H: Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. Journal of the American College of Cardiology 2017, 70(1).
2. Krishnamurthi R, Ikeda T, Feigin V: 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(2):171–179.
3. King D, Wittenberg R, Patel A, Quayyum Z, Berdunov V, Knapp M: The future incidence, prevalence and costs of stroke in the UK. Age and ageing 2020, 49(2):277–282.
4. Boehme AK, Esenwa C, Elkind MSV: Stroke Risk Factors, Genetics, and Prevention. Circulation Research 2017, 120(3):472–495.
5. Wajngarten M, Silva GS: Hypertension and Stroke: Update on Treatment. European Cardiology Review 2019, 14(2):111–115.
6. Du X, McNamee R, Cruickshank K: Stroke Risk from Multiple Risk Factors Combined with Hypertension: A Primary Care Based Case-control Study in a Defined Population of Northwest England. Annals of epidemiology 2000, 10(6):380–388.
7. Roy-O'Reilly M, McCullough LD: Age and Sex Are Critical Factors in Ischemic Stroke Pathology. Endocrinology 2018, 159(8):3120–3131.
8. Yao Q, Zhang J, Yan K, Zheng Q, Li Y, Zhang L, Wu C, Yang Y, Zhou M, Zhu C: Development and validation of a 2-year new-onset stroke risk prediction model for people over age 45 in China. Medicine 2020, 99(41):e22680–e22680.
9. Chun M, Clarke R, Cairns BJ, Clifton D, Bennett D, Chen Y, Guo Y, Pei P, Lv J, Yu C, Yang L, Li L, Chen Z, Zhu T: Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. Journal of the American Medical Informatics Association 2021, 28(8):1719–1727.
10. Wang Y, Deng Y, Tan Y, Zhou M, Jiang Y, Liu B: A comparison of random survival forest and Cox regression for prediction of mortality in patients with hemorrhagic stroke. BMC medical informatics and decision making 2023, 23(1):1–215.
11. Chen Y, Chung J, Yeh Y, Lou S, Lin H, Lin C, Hsien H, Hung K, Yeh SJ, Shi H: Predicting 30-Day Readmission for Stroke Using Machine Learning Algorithms: A Prospective Cohort Study. Frontiers in neurology 2022, 13:875491.
12. MacCarthy G, Pazoki R: Evaluation of Machine Learning and Traditional Statistical Models to Assess the Value of Stroke Genetic Liability for Prediction of Risk of Stroke Within the UK Biobank. Healthcare (Basel) 2025, 13(9):1003.
13. Papadopoulou A, Harding D, Slabaugh G, Marouli E, Deloukas P: Prediction of atrial fibrillation and stroke using machine learning models in UK Biobank. Heliyon 2024, 10(7):e28034.
14. Gong L, Chen S, Yang Y, Hu W, Cai J, Liu S, Zhao Y, Pei L, Ma J, Chen F: Designing machine learning for big data: A study to identify factors that increase the risk of ischemic stroke and prognosis in hypertensive patients. Digital health 2024, 10:20552076241288833.
15. Yang Y, Zheng J, Du Z, Li Y, Cai Y: Accurate Prediction of Stroke for Hypertensive Patients Based on Medical Big Data and Machine Learning Algorithms: Retrospective Study. JMIR medical informatics 2021, 9(11):e30277.
16. Li A, Ji Y, Zhu S, Hu Z, Xu X, Wang Y, Jian X: Risk probability and influencing factors of stroke in followed-up hypertension patients. BMC cardiovascular disorders 2022, 22(1):1–10.
17. MacCarthy G, Pazoki R: Using Machine Learning to Evaluate the Value of Genetic Liabilities in the Classification of Hypertension within the UK Biobank. Journal of clinical medicine 2024, 13(10):2955.
18. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS medicine 2015, 12(3):e1001779.
19. Bycroft C, Freeman C, Petkova D, Band G, Elliott L, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'connell J, Cortes A, Welsh S, Mcvean G, Leslie S, Donnelly P, Marchini J: Genome-wide genetic data on ~500,000 UK biobank participants. bioRxiv 2017, :.
20. Welsh S, Peakman T, Sheard S, Almond R: Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. BMC Genomics 2017, 18(1):26.
21. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J: The Uk biobank resource with deep phenotyping and genomic data. Nature 2018, 562(7726):203–209.

22. Sacks DB, Arnold M, Bakris GL, Bruns DE, Horvath AR, Kirkman MS, Lernmark A, Metzger BE, Nathan DM: Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. Clinical chemistry (Baltimore, Md.) 2011, 57(6):e1–e47.
23. Malik R, Rannikmäe K, Traylor M, Georgakis MK, Sargurupremraj M, Markus HS, Hopewell JC, Debette S, Sudlow CLM, Dichgans M: Genome-wide meta-analysis identifies 3 novel loci associated with stroke. Annals of neurology 2018, 84(6):934–939.
24. Suresh K, Severn C, Ghosh D: Survival prediction models: an introduction to discrete-time modeling. BMC medical research methodology 2022, 22(1):1–207.
25. Lunardon N, Menardi G, Torelli N: ROSE: a Package for Binary Imbalanced Learning. The R journal 2014, 6(1):79.
26. Wei Q, Dunbrack J, Roland L: The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. PLoS ONE 2013, 8(7):e67863.
27. Greenwood CJ, Youssef GJ, Letcher P, Macdonald JA, Hagg LJ, Sanson A, Mcintosh J, Hutchinson DM, Toumbourou JW, Fuller-Tyszkiewicz M, Olsson CA: A comparison of penalised regression methods for informing the selection of predictive markers. PloS one 2020, 15(11):e0242730.
28. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ: Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Statistics in medicine 2016, 35(7):1159–1177.
29. PLATT J: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Largin Margin Classifiers 2000, .
30. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L: A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association 2020, 27(4):621–633.
31. Suresh K, Severn C, Ghosh D: Survival prediction models: an introduction to discrete-time modeling. BMC medical research methodology 2022, 22(1):1–18.
32. Ding Q, Liu S, Yao Y, Liu H, Cai T, Han L: Global, Regional, and National Burden of Ischemic Stroke, 1990–2019. Neurology 2022, 98(3):e279–e290.
33. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, Arnold M, Bell S, Bolton T, Burgess S, Dudbridge F, Guo Q, Sofianopoulou E, Stevens D, Thompson JR, Butterworth AS, Wood A, Danesh J, Samani NJ, Inouye M, Di Angelantonio E: Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. PLoS medicine 2021, 18(1):e1003498.

Supplementary Material for Chapter 5

Chapter Five: Using discrete- and continuous-time machine learning models (Nnet, CoxNet, GLMnet) to explore sex and age differences in stroke prediction among hypertensive individuals.

Supplementary Table 1: **Conventional Risk Factor and Genetic Risk Factor**

Conventional risk factors
Age
Sex
Body mass index (BMI**
Diabetes Mellitus (DM)
Total Cholesterol (TC)
Low lipoprotein (LDL)**
Alcohol
Smoking
Genetic risk factor
Stroke genetic liability (SGL)

** Factors are not significant and not included in the prediction models

Supplementary Table 2: Tuning Techniques for Hyperparameter for Each Machine Learning

Machine Learning	**Packages**	**Function**	**Hyperparameters**	**Optimal tuning parameters for Models**
CoxPH	survival	coxph	N/A	N/A
Penalized Cox model (CoxNet)	glmnet	cv.glmnet with 10-fold	alpha and lambda	alpha = 0.5 lambda = 0.0001
Penalised logistic model (GLMnet)	glmnet	cv.glmnet with 10-fold	alpha and lambda	alpha = 0.5, lambda = 0.002.
Neural network (Nnet)	nnet	caret with 10-fold	size and decay	size = 5 decay = 0.1

Supplementary Table 3: Assessment of the Proportional Hazard (PH) Assumption Using the Global Schoenfield Test

Variables	Chisq	df	P-value
Age	0.04	1	0.85
Sex	0.12	1	0.73
Diabetes Mellitus (DM)	0.38	1	0.54
Total Cholesterol (TC)	1.05	1	0.31
Alcohol	0.14	2	0.93
Smoking	1.97	2	0.37
Stroke Genetic liability (SGL)	0.43	1	0.51
GLOBAL	3.99	9	0.91

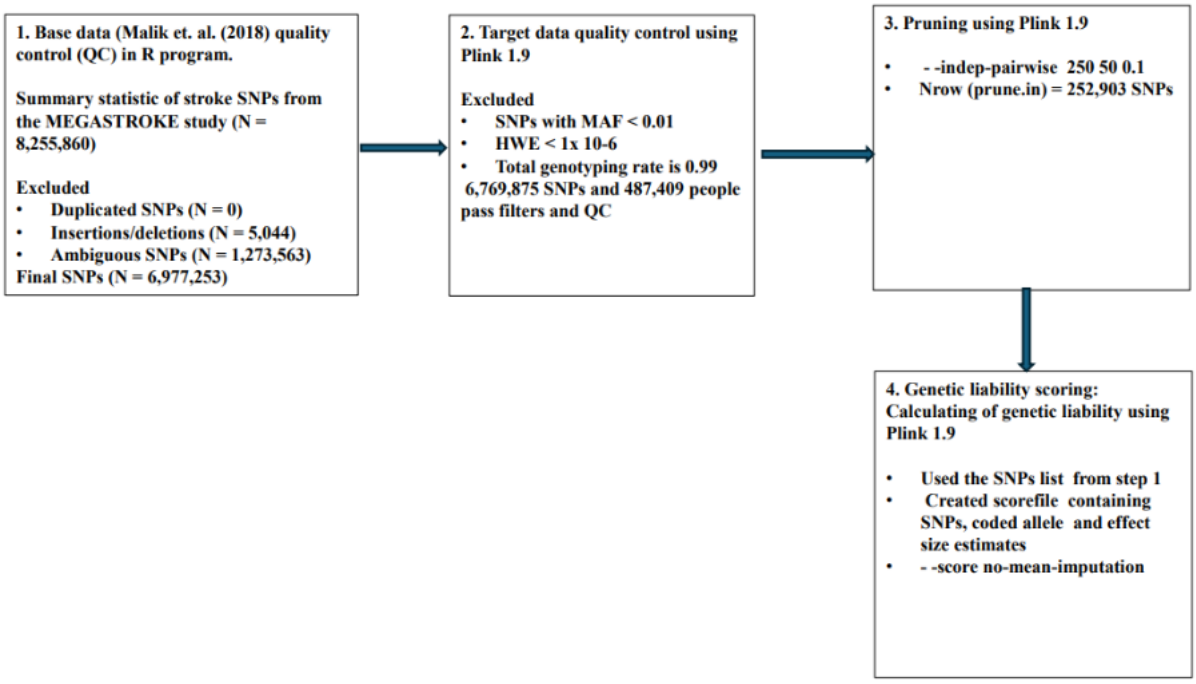
Supplementary Table 4: The Results of the Model Performance for Incident Stroke Prediction in the Hypertensive Patients in the UK Biobank (Without Genetic liability).

Models	Model Features	Sample	AUC (95% CI)	Brier Score
CoxPH				
	Conventional factors	whole population (n =116,216)	67.0 (64, 70)	0.01
	Conventional factors	Men (n = 55,269)	67.0 (64, 70)	0.01
	Conventional factors	Women (n = 60,947)	64.0 (60, 68)	0.01
	Conventional factors	Age > 59 (n = 55370)	60.0 (56, 64)	0.02
	Conventional factors	Age < 59 (n = 60846)	58.0 (53, 63)	0.02
CoxNet				
	Conventional factors	whole population (n=116,216)	67.0 (64, 70)	0.01
	Conventional factors	Men (n = 55,269)	67.0 (64, 71)	0.01
	Conventional factors	Women (n = 60,947)	65.0 (60, 69)	0.01
	Conventional factors	Age > 59 (n = 55370)	60.0 (57, 63)	0.02
	Conventional factors	Age < 59 (n = 60846)	56.0 (51, 61)	0.01

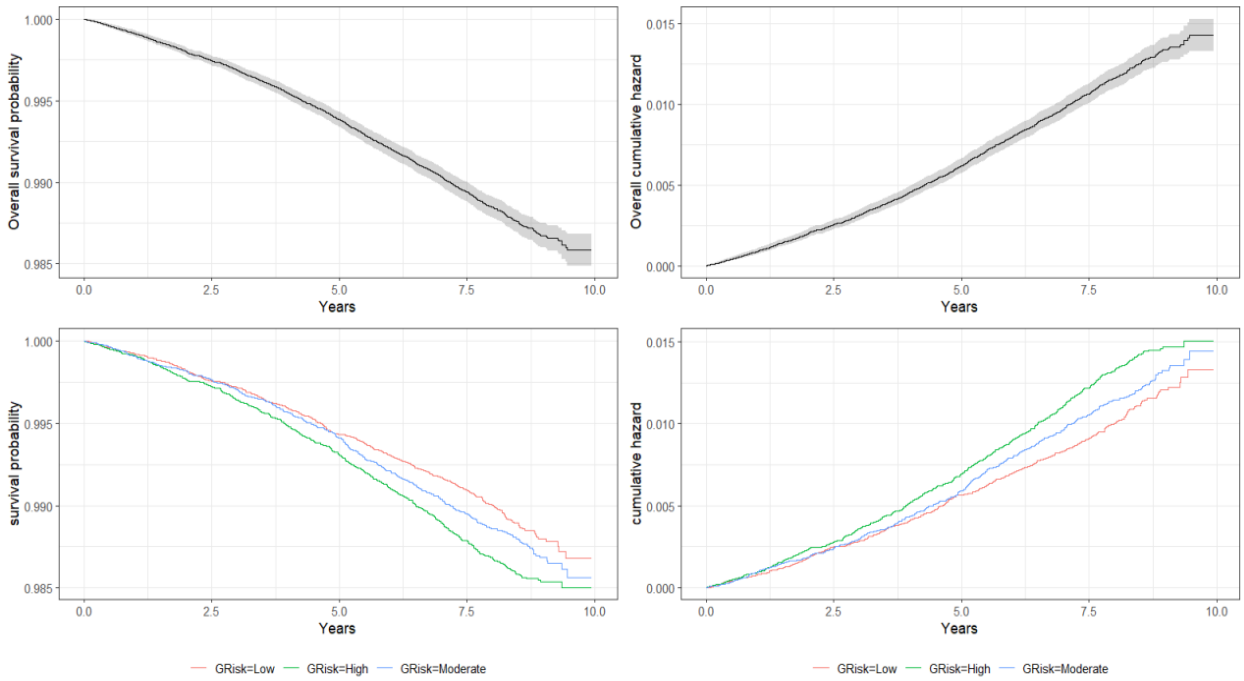
Supplementary Table 4 Continued

Models	Model Features	Sample	AUC (95% CI)	Brier Score
GLMnet				
	Conventional factors	whole population (n=116,216)	67.0 (64, 70)	0.001
	Conventional factors	Men (n = 55,269)	67.0 (63, 70)	0.002
	Conventional factors	Women (n = 60,947)	65.0 (61, 69)	0.001
	Conventional factors	Age > 59 (n = 55370)	61.0 (58, 65)	0.002
	Conventional factors	Age < 59 (n = 60846)	57.0 (52, 62)	0.001
Nnet				
	Conventional factors	whole population (n=116,216)	65.0 (63, 69)	0.001
	Conventional factors	Men (n = 55,269)	65.0 (62, 69)	0.002
	Conventional factors	Women (n = 60,947)	64.0 (59, 68)	0.001
	Conventional factors	Age > 59 (n = 55370)	60.0 (56, 63)	0.002
	Conventional factors	Age < 59 (n = 60846)	55.0 (50, 60)	0.001

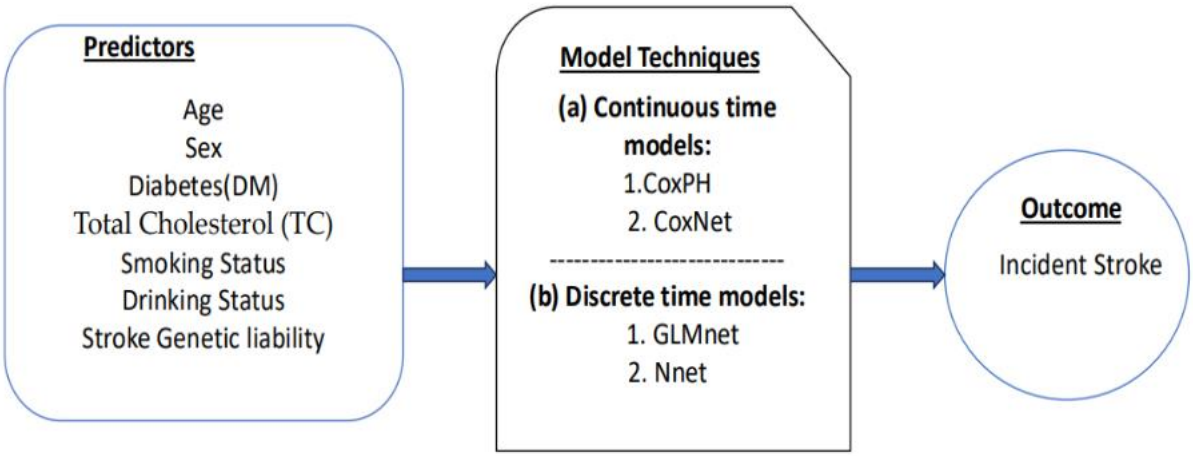
Supplementary Figure 1



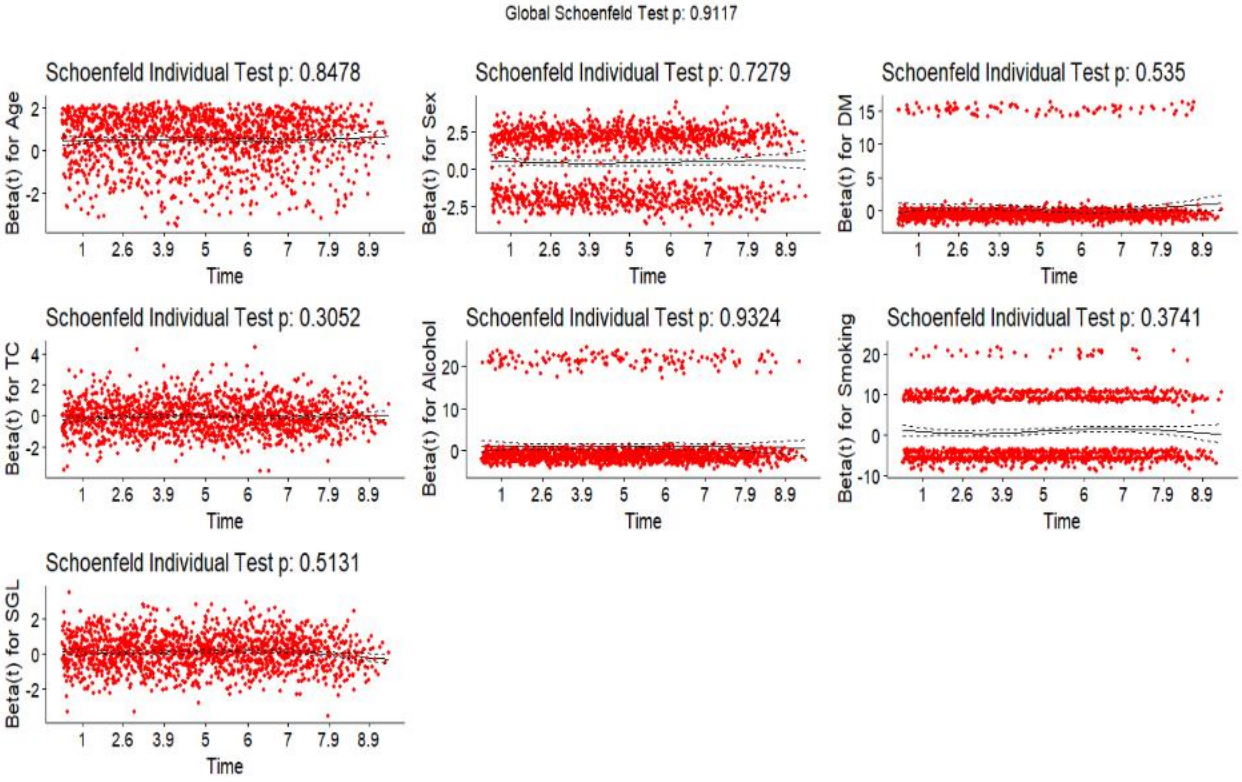
Supplementary Figure 2



Supplementary Figure 3

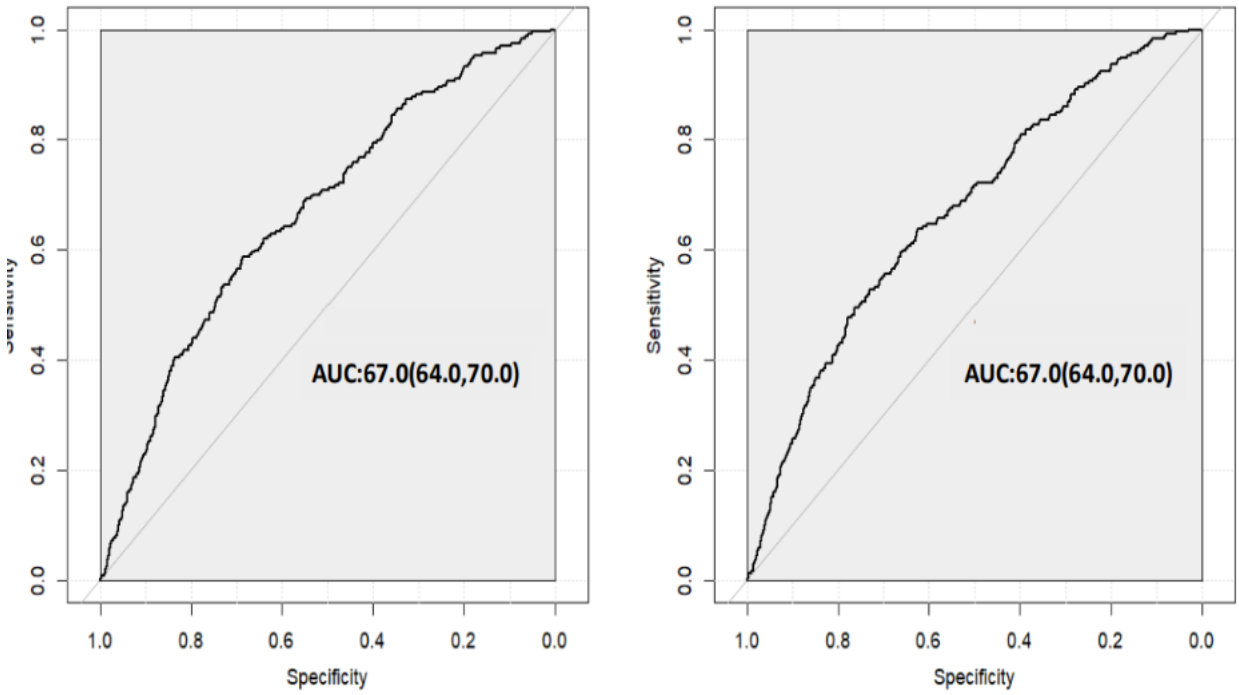


Supplementary Figure 4



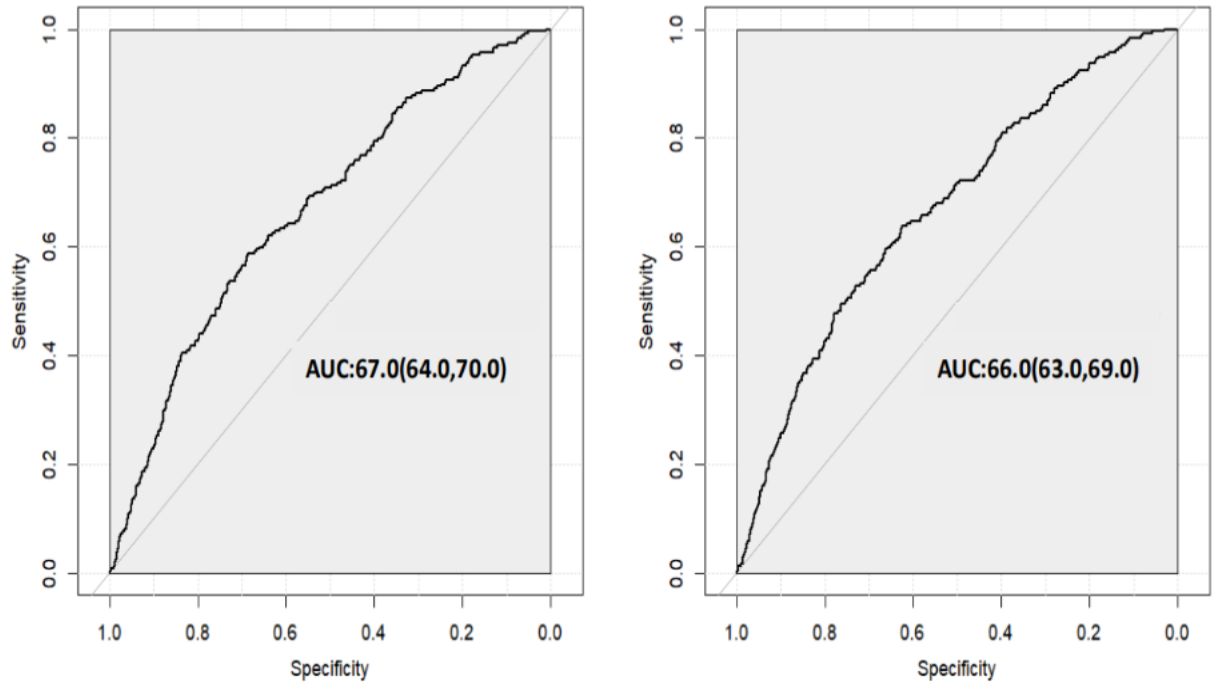
Supplementary Figure 4: Assessment of the proportional hazard assumption using Schoenfeld residuals. The global test revealed no significant deviation from proportionality ($p > 0.05$), confirming the model's validity.

Supplementary Figure 5



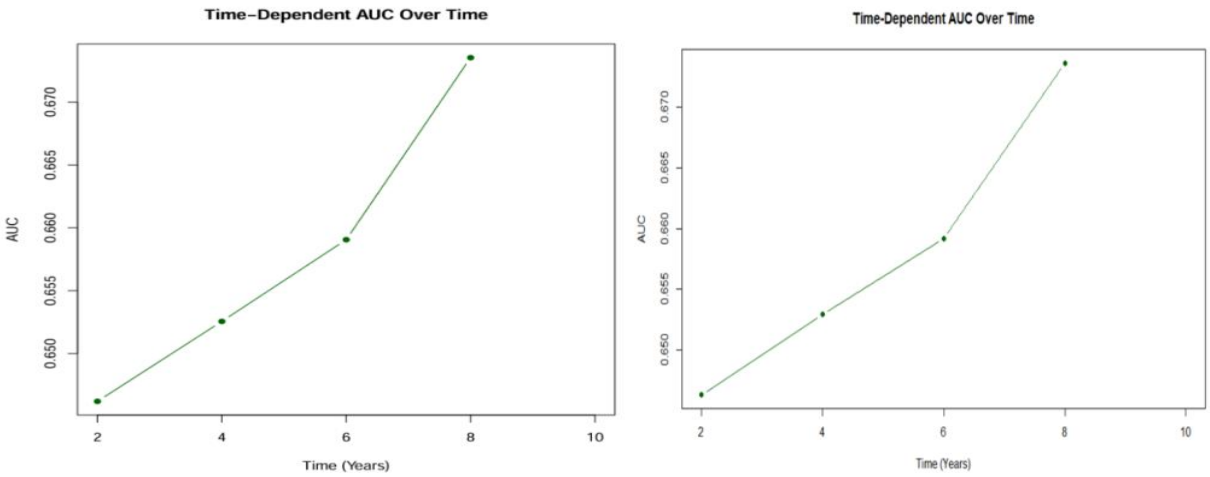
Supplementary Figure 5: ROC-AUC curve for CoxPH (Left panel). ROC-AUC curve for CoxNet(Right panel|)

Supplementary Figure 6



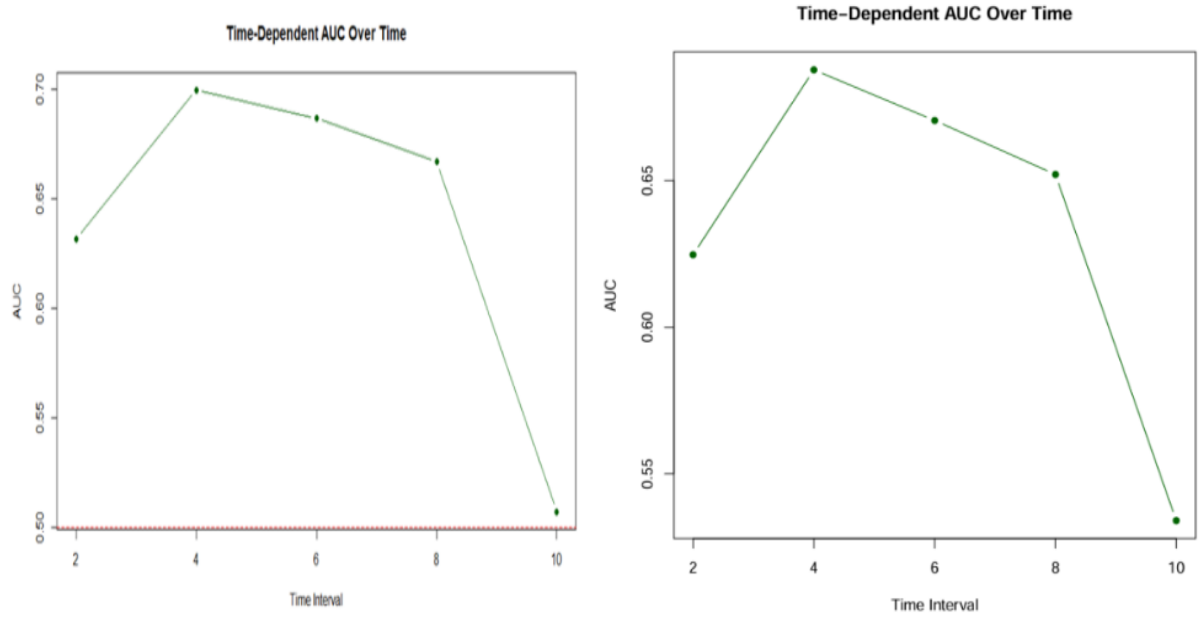
Supplementary Figure 6: ROC-AUC curve for GLMnet (Left panel). ROC-AUC curve for Nnet (Right panel|)

Supplementary Figure 7



Supplementary Figure 7: ROC-AUC curve CoxPH (Left panel). ROC-AUC curve for CoxNet(Right panel|)

Supplementary Figure 8



Supplementary Figure 8: ROC-AUC curve GLMnet (Left panel). ROC-AUC curve for Nnet (Right panel|)

6 CHAPTER SIX. GENERAL DISCUSSION

6.1 Overview of Aims and Findings

This thesis utilized both conventional statistical and machine learning methods to explore the influence of genetic liabilities for hypertension classification and prediction of stroke risks in participants of European ancestry within the UK Biobank. In three related studies, the main aim was to examine whether or not stroke genetic liability and genetic liability to CVD risk factors can improve disease classification and prediction when combined with demographic, lifestyle, and clinical risk factors. The second aim was to examine the relative strength of performance of modelling methods, and the third aim was to examine whether predictive ability varied across subpopulations grouped by age and sex.

This thesis examines a key question not yet fully answered in the current medical literature: How much additional predictive ability can genetic liability contribute to the classification of hypertension and the prediction of stroke risk when included in traditional risk factor models, principally using machine learning models? Existing literature [1-3] provides evidence that genetic liability can slightly improve discrimination, calibration, and risk reclassification when added to conventional risk factor models. Nevertheless, gaps remain, largely in part, many studies have not fully examined genetic liability data in the prediction and classification of complex diseases such as stroke and hypertension. There are also unanswered questions about subgroups, the magnitude of improvement, and which specific prediction model methods yield the greatest incremental value from genetic liability.

The findings from this research have expanded our understanding in many important ways.

First, this research has demonstrated that genetic liability can improve the model's ability to predict disease. Genetic liability to a disease can be quantified at birth, even before clinical symptoms manifest, thus potentially creating an important window for possible early intervention. Second, this research has shown how much the key metrics for ascertaining a model's performance (AUC, calibration, and reclassification) improve when genetic liability is included in the prediction models. Also, the improvements remained consistent across the various subgroups, which is critical in building reliable and robust prediction tools.

Third, the research has highlighted the nature of interactions between genetic liability and other conventional risk factors. For instance, through correlation analysis, we have demonstrated that stroke genetic liability has an independent relationship with the established risk factors. That means genetic liability provides an independent and unique contribution to disease risk.

Fourth, by using and comparing machine learning models to more traditional statistical approaches, this research provides practical trade-offs for modeling techniques. For instance, the study considered the balance between a model's complexity and the added value it offers, while also considering the importance of the interpretability of overall results.

Utilizing genetic information and traditional risk factors with machine learning methods to predict cardiovascular disease risk is a promising but relatively new and unexplored area. As systematic reviews suggest, a growing number of studies are examining how artificial intelligence and machine learning algorithms could optimize genetic liability (AI-optimized PRS) to improve cardiovascular disease prediction [4-6]. For example, Drouard and colleagues [7] have developed machine learning techniques to predict CVD risk factors from multi-omics data while exploring different data type combinations or by integrating multiple data types such as clinical data, genetic data, and biomarker data, within the combined model. However, there are still fewer or no studies that have explored the contribution of stroke genetic liability to stroke risk prediction in subgroups such as age and sex analysis.

Thus, this research contributes methodologically to ongoing research about the utility of genetic information in disease risk prediction, in demonstrating how to evaluate machine learning and genetic liability, and also by illustrating which subgroup the genetic liability contributes most to the predictive accuracy. Therefore, this research helps serve as a bridge that supports the pathway towards more personalized and effective risk prediction.

6.1.1 Genetic Liabilities and Hypertension Prediction

The first study explored the application of machine learning techniques (Random Forest and Neural Network) to assess the predictive value of a combination of genetic liabilities for type 2 diabetes, obesity traits, lipid traits, and smoking traits in a single model for hypertension classification. Though genetic liabilities capture an inherited risk of the disease, this research demonstrated that their contribution to classification performance is

modest but significant when they are used in combination with existing clinical and lifestyle risk factors.

6.1.2 Stroke Genetic Liability and Stroke Prediction

The second study focused on stroke, examining whether machine learning methods such as Random Forest, Decision Tree, and Gradient Boost Machine could improve stroke risk prediction compared to the Cox proportional hazards model when stroke genetic liability was included. In this study, similar results to those in the first study were observed. While stroke genetic liability was statistically significant, its contribution to risk prediction was modest. Importantly, the study found that the Cox proportional hazards model and machine learning models often produced similar prediction performance.

6.1.3 Stroke Prediction Among Hypertensive Individuals

The analysis of the third study focused on hypertensive patients, a high-risk group for stroke. The study compared the prediction performance of continuous-time survival models (Cox proportional hazards model, Penalized Cox model) to discrete-time survival models (Penalised logistic model and Neural Network) and assessed whether prediction performance differed by age and sex.
There was significant variability in results, and the study found that both the predictive performance of models and the contributions of stroke genetic liability differed across subgroups. For example, comparing the predictive value of genetic liability by sex and age group, it was observed that the genetic liability of stroke had modestly improved the model's prediction performance, particularly for older participants and hypertensive men. The improvement may not be statistically significant as the confidence interval overlapped. Notably, the Cox proportional hazards model showed superior prediction performance compared to machine learning models in subgroup analyses. This illustrates the efficacy of the Cox proportional regression model, especially when dealing with time-to-event epidemiological data. Although the continuous-time survival models may provide more clinical understanding of various risk pathways, discrete-time survival models offer better interpretability and allow flexibility when working with censored outcomes.

6.2 Broader Implications of the Three Studies

In all three studies, all genetic liabilities included were significantly associated with the outcome. However, the additional predictive value gained from the

genetic liabilities was modest. This finding is consistent with the wider literature on genetic liability, which proves that while genetic liability can capture important inherited risk, the established risk factors remain the dominant and important predictors for complex diseases such as hypertension and stroke.

The machine learning versus traditional statistical methods comparison provided an interesting perspective on methodology. Here, it is observed that machine learning models underperformed compared to the Cox proportional hazards model when using a well-defined epidemiological dataset. Empirically, machine learning models outperform traditional statistical models in the context of high-dimensional or unstructured data such as multi-omics data or electronic health records. For the epidemiology of cardiovascular disease, traditional regression modeling is still a valuable technique that is easy to understand and interpret in the clinical context.

The subgroup analysis in predicting stroke further provides validation that prediction performance varies by subgroup. Specifically, stroke genetic liability improved stroke prediction in older than younger hypertensive patients. In addition, it has also improved stroke prediction in hypertensive men compared to hypertensive women. Stroke epidemiology research has established the evidence that there exist sex and age differences in stroke risk [8], which underscores the need for sex-specific risk models.

From an applicable translational approach, these findings indicate that although genetic information is unlikely to fully replace traditional clinical risk factors soon, it could facilitate early-life risk stratification or help in risk prediction for high-risk groups. Moreover, utilizing machine learning methods may be especially useful in discovering complex interactions between predictors, which may assist in future research regarding disease mechanisms or causes. Finally, this research emphasizes the importance of large-scale biobank resources in general. The biobanks' resources are critical in systematically assessing genetic and non-genetic predictors for a more expansive application of the modeling framework and approaches.

6.2.1 Strengths

The overall strengths of this thesis were the use of a well-characterised cohort study with a large sample size. Likewise, we also systematically compared multiple modeling approaches and innovatively assessed how the risk of stroke

predictions varies by group representation. Another strength was the inclusion of multiple genetic liabilities in one single prediction model to identify the best performing classification model for hypertension. A distinctive feature and the strength of the second and third studies compared with previous studies is that we generated genetic liability for stroke using over 250,000 genetic variants.

More of the strengths have been stated in the three related studies.

6.2.2 Limitations

This study has several limitations that should be addressed. First, the analyses were limited to participants of European ancestry within the UK Biobank, and the whole-genome genetic liabilities were primarily calculated from GWAS comprised of European ancestry. These restricts the generalisability of our findings [9]. Second, the analyses in this study were further restricted to participants who have been identified as or diagnosed with hypertension. These factors might limit the generalisation of our findings to a wider UK population from different ancestries or health statuses. This means that the predictive performance observed in this research cannot be extended to other ancestry groups, or to participant with different socioeconomic or health statuses. Future research should be expanded to include participants with other ancestries and socioeconomic strata.

Third, the measure of genetic risk is limited to genetic liability calculated from genome-wide summary statistics. Although the genetic liability can explain a large proportion of common variant predisposition, the genetic liability calculations do not include variants that are rare or gene-gene and gene-environment interactions that may also contribute to CVD risk [10]. An alternative method, such as a machine learning driven SNP selection method to generate genetic liability, could have been employed. Third, there is the possibility of healthy volunteer bias in the UK Biobank cohort because self-reported lifestyle factors may be different within the UK Biobank.

6.2.3 Future Directions

Future research should examine the integration of multi-omics (e.g., transcriptomics, metabolomics, and proteomics) with genetic liability data in prediction models. This could provide a more general risk profile.

6.2.4 Conclusion

This thesis ultimately illustrates that genetic liability provides modest but consistent improvement to the prediction of both outcomes. This was observed with stroke prediction, particularly in subgroups such as older and male hypertensive individuals. Although it has been demonstrated in many studies that the machine-learning approaches are powerful, the traditional Cox regression employed in this research has consistently outperformed any machine learning in this specific context. Overall, our findings support and underscore that genetic liability should be considered as a valuable complement to legacy clinical and lifestyle risk factors, rather than their replacement.

The addition of stroke genetic liability in the models improved stroke prediction for a small percentage of the population. Therefore, its utilisation and efficacy in clinical practice is uncertain. The conventional risk factors may still have more influence on the prediction of stroke in hypertensive patients. The findings suggest that genetic liability alone has limited predictive value for most people, but it might still have a role in highly targeted interventions. In terms of cost effectiveness, given that only a small percentage of the population benefits from genetic risk scores, in future health economics studies are needed to establish if the costs of genetic testing could outweigh the potential improvements in stroke prevention especially in risk group such as hypertensive patients.

References

1. Chung R, Xu Z, Arnold M, Ip S, Harrison H, Barrett J, Pennells L, Kim LG, Di Angelantonio E, Paige E, Ritchie SC, Inouye M, Usher-Smith JA, Wood AM: **Using Polygenic Risk Scores for Prioritizing Individuals at Greatest Need of a Cardiovascular Disease Risk Assessment.** Journal of the American Heart Association 2023, **12**(15):e029296.

2. Li L, Pang S, Starnecker F, Mueller-Myhsok B, Schunkert H: **Integration of a polygenic score into guideline-recommended prediction of cardiovascular disease.** European Heart Journal 2024, .

3. Sun L, Pennells L, Kaptoge S, Nelson CP, Ritchie SC, Abraham G, Arnold M, Bell S, Bolton T, Burgess S, Dudbridge F, Guo Q, Sofianopoulou E, Stevens D, Thompson JR, Butterworth AS, Wood A, Danesh J, Samani NJ, Inouye M, Di Angelantonio E: **Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses.** PLoS medicine 2021, **18**(1):e1003498.

4. Alireza Z, Maleeha M, Kaikkonen M, Fortino V: **Enhancing prediction accuracy of coronary artery disease through machine learning-driven genomic variant selection.** J Transl Med 2024, **22**(1):356–14.

5. Hosseini K, Anaraki N, Dastjerdi P, Kazemian S, Hasanzad M, Alkhouli M, Alam M, Nasir K, Rana JS, Bhatt AB: **Bridging Genomics to Cardiology Clinical Practice.** JACC. Advances (Online) 2025, **4**(6):101803.

6. Khanna NN, Singh M, Maindarkar M, Kumar A, Johri AM, Mentella L, Laird JR, Paraskevas KI, Ruzsa Z, Singh N, Kalra MK, Fernandes JFE, Chaturvedi S, Nicolaides A, Rathore V, Singh I, Teji JS, Al-Maini M, Isenovic ER, Viswanathan V, Khanna P, Fouda MM, Saba L, Suri JS: **Polygenic Risk Score for Cardiovascular Diseases in Artificial Intelligence Paradigm: A Review.** Journal of Korean medical science 2023, **38**(46):e395.

7. Drouard G, Mykkänen J, Heiskanen J, Pohjonen J, Ruohonen S, Pahkala K, Lehtimäki T, Wang X, Ollikainen M, Ripatti S, Pirinen M, Raitakari O, Kaprio J: **Exploring machine learning strategies for predicting cardiovascular disease risk factors from multi-omic data.** BMC Med Inform Decis Mak 2024, **24**(1):116–18.

8. Roy-O'Reilly M, McCullough LD: **Age and Sex Are Critical Factors in Ischemic Stroke Pathology.** Endocrinology 2018, **159**(8):3120–3131.

9. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ: **Clinical use of current polygenic risk scores may exacerbate health disparities.** Nat Genet 2019, **51**(4):584–591.

10. Choi KW, Stein MB, Nishimi KM, Ge T, Coleman JRI, Chen C, Ratanatharathorn A, Zheutlin AB, Dunn EC, Breen G, Koenen KC, Smoller JW: **An Exposure-Wide and Mendelian Randomization Approach to Identifying Modifiable Factors for the Prevention of Depression.** The American journal of psychiatry 2020, **177**(10):944–954.