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.

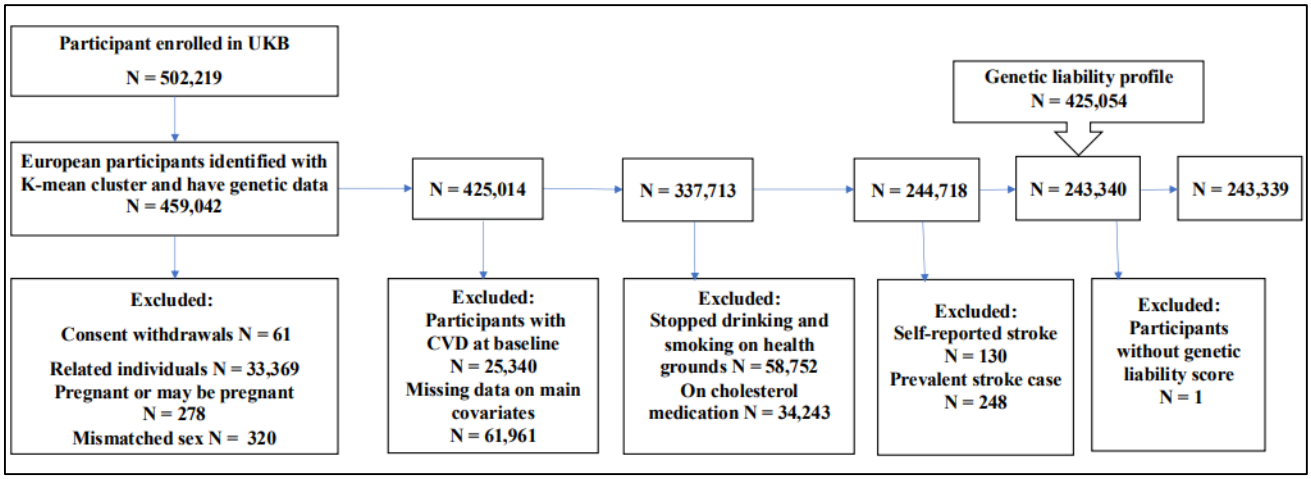

Figure 1. Exclusion criteria of the study: The flowchart for selecting research participants. At the start of this study, the UK Biobank (UKB) had over 500,000 participants. We employed the K-means cluster approach to extract 459,042 European-ancestry subjects. The final dataset had 243,339 people who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

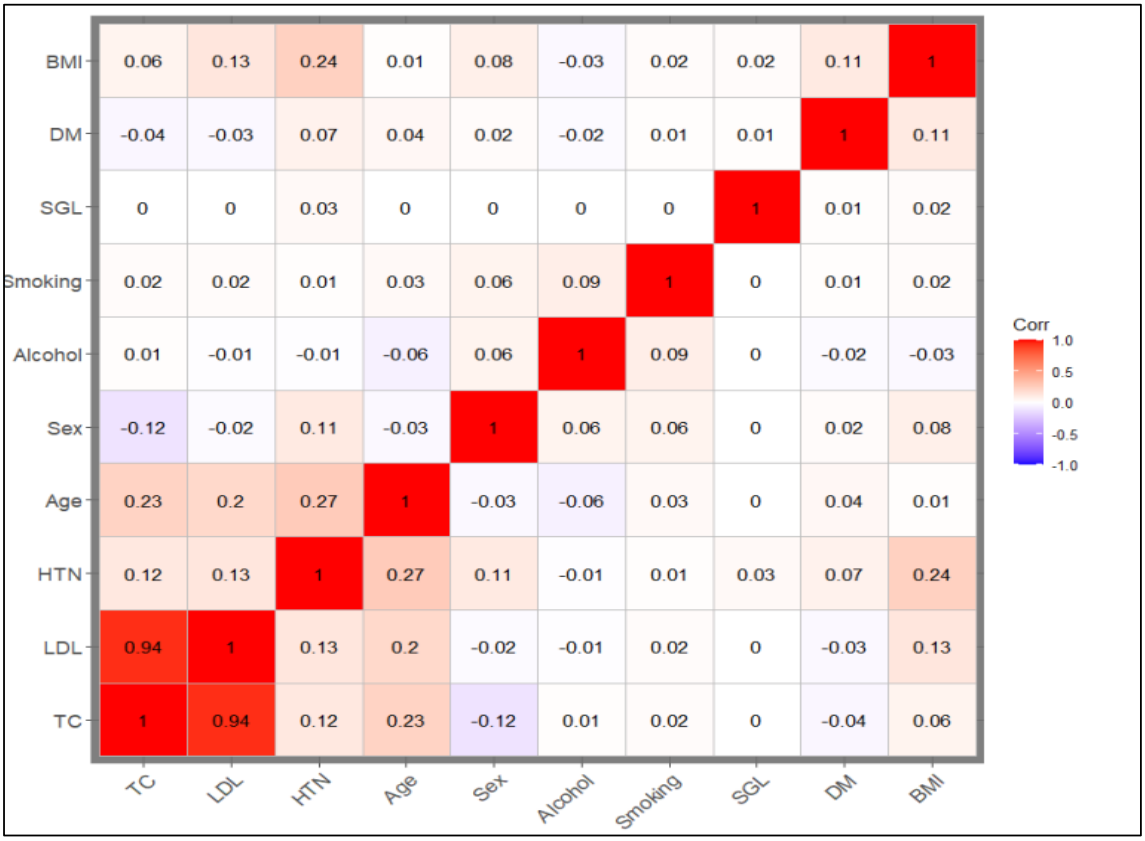

Figure 2. Correlation matrix plot: The plot shows the correlation coefficients between numerical features. TC and LDL are highly correlated ($r^2 > 0.8$). TC was excluded from further analysis (prediction model construction). BMI: body mass index; TC: total cholesterol; LDL: low-density lipoprotein cholesterol;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

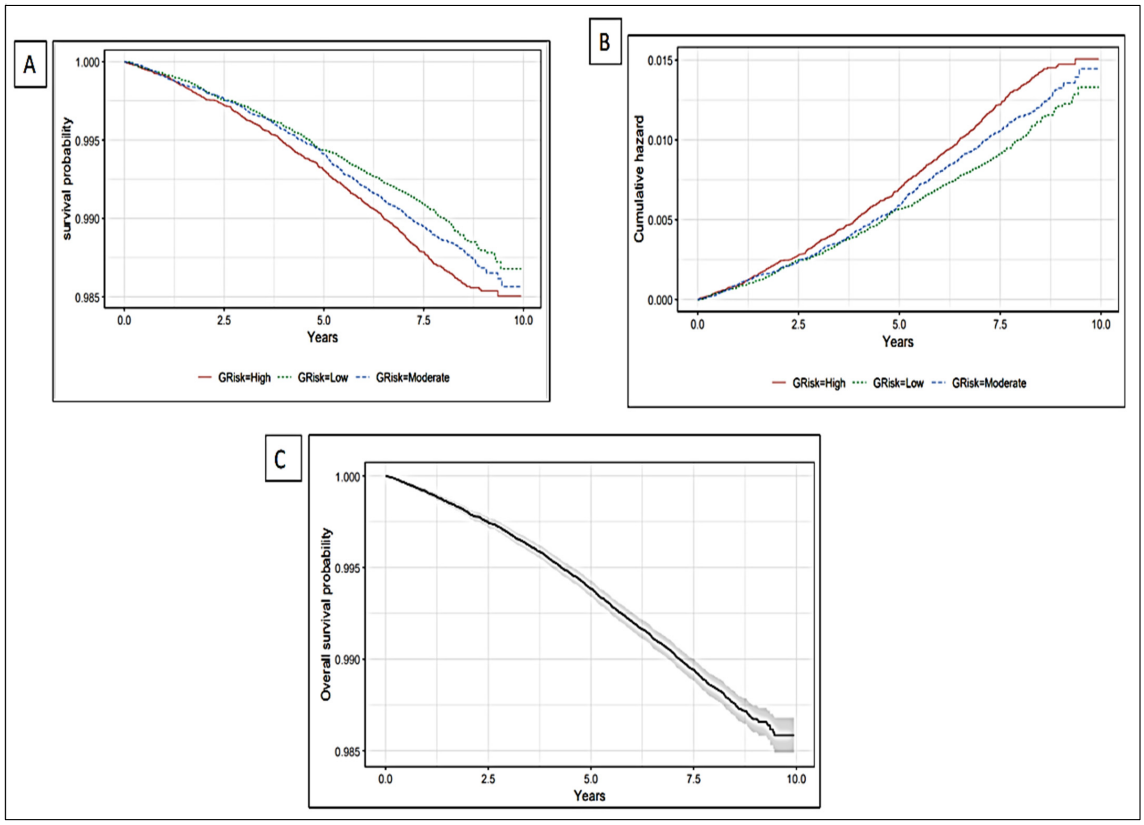

Figure 3. Survival probability and cumulative hazard plot stratified by genetic risk level: (**A**) Survival probability plot stratified by genetic risk level. (**B**) Cumulative hazard plot stratified by genetic risk level. (**C**) Overall survival probability of the study population. (**A**,**B**) 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.

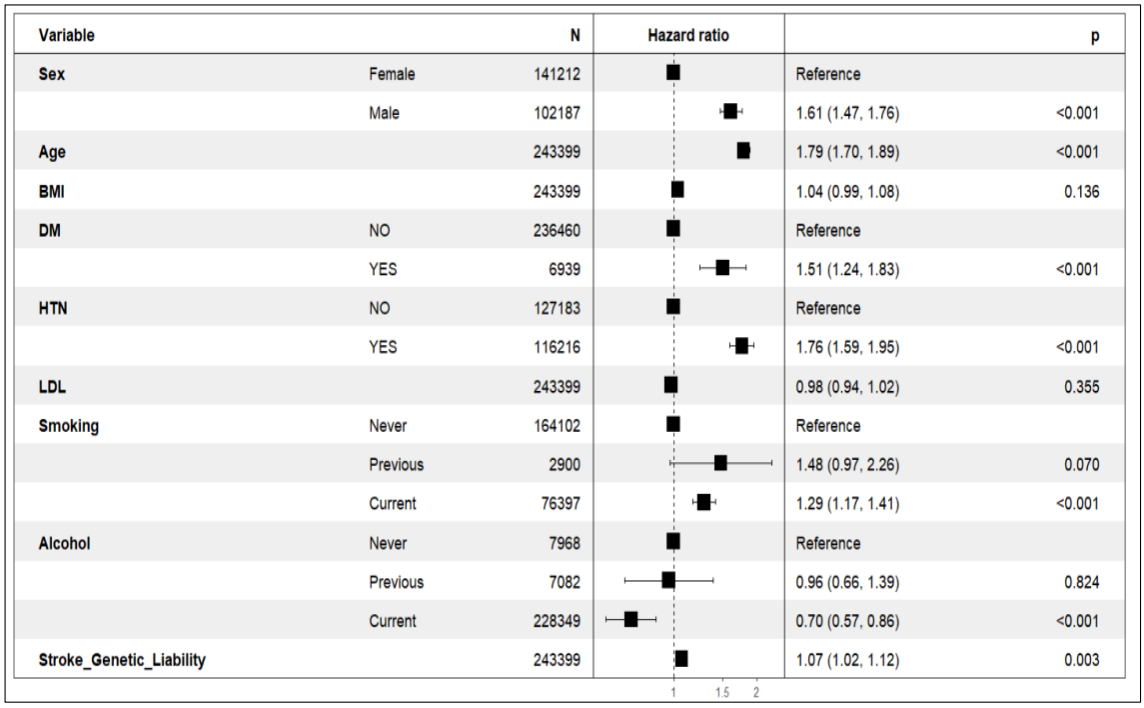

Figure 4. Forest plot of the full Cox proportional hazard model: The vertical line at the hazard ratio (HR) = 1 is the reference line. The horizontal line represents the confidence interval (CI). HTN: hypertension.

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]