scientific reports

Check for updates

OPEN

Data-driven prediction of cardiovascular and cerebrovascular diseases in a nationwide study

Sehyun Kim[1], Beomsang Ryu[2], Mingee Choi[3], Sangyon Lee[1], Jaeyong Shin[3]✉ & Sok Chul Hong[1]✉

As the importance of the prevention and premanagement of cardiovascular and cerebrovascular diseases continues to emerge, research is being conducted globally to create and compare risk factor prediction models using health examination big data. In this study, health insurance data were used to predict the incidence of cardiocerebrovascular disease using various models and compare the performance of the models on samples with different initial risk levels. This study analyzed data from 410,859 individuals from the National Health Insurance Service between 2002 and 2019. This study deployed various linear models to predict the occurrence of cardiocerebrovascular diseases in two distinct samples. Models based on logistic regression analysis with penalty terms on the objective function were used, and their predictive performances were compared using multiple evaluation metrics, including the area under the receiver operating characteristic curve. The logistic regression model incorporating variables selected by the LASSO algorithm exhibited superior predictive performance relative to other models, although the differences were not statistically significant. The models demonstrated improved performance for samples with higher incidence rates and initial risk levels. This study predicted and compared the incidence of cardiocerebrovascular disease (CCVD) in patients with different health conditions using national sample cohort data from the National Health Insurance Service. The findings underscore the importance of developing diverse models to predict diseases like CCVD, which have high medical costs and incidence rates, thus informing the development of healthcare policies.

Keywords Cardiovascular system, Risk prediction, Big data, Health analytics

Abbreviations

AUROC	Area under the receiver operating characteristic curve
BMI	Body mass index
CCVD	Cardiocerebrovascular disease
DALY	Disability-adjusted life-years
FN	False negative
FP	False positive
NHIS-NSC	National Health Insurance Service National Sample Cohort
SBP	Systolic blood pressure
TN	True negative
TP	True positives
IHD	Ischemic heart disease
AMI	Acute myocardial infarction

Due to an aging population and shifts in health behaviors, the incidence of cardiovascular diseases is steadily rising, contributing to an increasing burden in terms of both disease prevalence and medical expenses[1–3]. In Korea, the mortality rate from cardiovascular and cerebrovascular diseases has risen by 7% over the past decade, with medical related to these conditions accounting for approximately 17% of the nation's total healthcare costs[4].

[1]Department of Economics, Seoul National University College of Social Science, Gwanak-gu, Seoul 08826, Republic of Korea. [2]Wellxecon, Gangnam-gu, Seoul, Republic of Korea. [3]Department of Preventive Medicine, Yonsei University College of Medicine, 50 Yonsei-ro, Seodaemun- gu, Seoul 03722, Republic of Korea. ✉email: drshin@yuhs.ac; sokchul.hong@snu.ac.kr

Maintaining a healthy lifestyle is crucial for preventing cardiovascular and cerebrovascular diseases. Risk factors, including smoking, unhealthy eating habits, physical inactivity, and alcohol consumption, have been consistently linked to an increased likelihood of developing cardiovascular conditions[5]. As individuals age, prolonged exposure to these harmful lifestyle behaviors, as well as environmental factors, further elevates the risk of cardiovascular and cerebrovascular diseases. Consequently, as the population ages, both the incidence of these diseases and the associated healthcare costs are expected to rise significantly[6].

Lifestyle habits related to cardiovascular and cerebrovascular diseases are modifiable, and these diseases can often be prevented through behavioral improvement[5]. Prevention and early management of cardiovascular and cerebrovascular are cost-effective strategies for reducing future disease burdens and mitigating soaring healthcare costs. Based on recent studies that utilize big data from both domestic and international healthcare systems, the evidence supporting the prevention of chronic diseases through healthcare interventions and behavioral changes has strengthened[6].

Previous studies have primarily focused on utilizing large-scale health examination data to identify risk factors for cardiovascular diseases and develop predictive models[7–10]. However, most of these studies have relied on simple regression analyses or machine learning techniques, which pose limitations due to increased complexity in interpretation and issues related to overfitting. To address these challenges, this study employs linear models and penalized linear models, specifically LASSO and Ridge regression, to predict the risk of cardiocerebrovascular disease incidence. These models enhance interpretability through variable selection and dimensionality reduction while mitigating overfitting and improving generalizability. Furthermore, to account for the heterogeneity in individual health status and risk factors, this study incorporates underwriting criteria commonly used in private health insurance to classify the sample into two groups: the standard risk group and the simplified risk group, based on initial risk levels. This classification facilitates personalized risk prediction and enables the development of differentiated risk management strategies, thereby enhancing the practical applicability of the findings.

Accordingly, the objectives of this study are as follows. First, using NHIS big data, this study aims to predict the incidence risk of cardiocerebrovascular diseases by applying various linear and penalized linear models. Second, it seeks to identify key risk factors for each group categorized based on initial risk levels. By achieving these objectives, this study aims to provide scientific evidence to support the development of more effective strategies for the prevention and management of cardiocerebrovascular diseases.

In this study, linear and penalized linear models were utilized to predict the incidence of cardiocerebrovascular diseases. While various predictive models for major chronic diseases exist, ranging from linear models to machine learning approaches, many lack external validation and are highly specific to the study context, which limits their generalizability. Therefore, using a nationally representative sample and methodologies commonly employed by insurers for risk assessment and underwriting, this study applies logistic regression models due to their widespread use, interpretability, and computational efficiency.

Methods

Data source and study population

This study utilized the National Health Insurance Service National Sample Cohort (NHIS-NSC) as the primary data source to develop predictive models for CCVDs (NHIS-2022-2-318, IRB No. P01-202206-01-031). The NHIS covers over 97% of the population, and a random sample of 2% was extracted using the proportional allocation method, taking into account factors such as sex, age, enrollment type, insurance premium quantile, and region. Sampling was based on national health insurance enrollees in 2006, and data were collected from various databases, including eligibility factors (sex, age, insurance premium, etc.), health checkup information (body mass index [BMI], waist circumference, blood pressure, self-reported questionnaires, etc.), and hospital utilization data (hospital admissions, diagnoses, prescriptions, etc.) from 2002 to 2019. Additional details regarding the representativeness of the NHIS-NSC and supplementary information can be found in the available sources[11] (Supplementary material).

Ascertainment of cardiovascular diseases

In this analysis, cerebrovascular disease (I60–I69) and ischemic heart disease (I20–I25) were used as dependent variables representing CCVDs. Subtypes of CCVDs were defined for additional analysis, including stroke (I60–I66, excluding I64), cerebral hemorrhage (I60–I62), and acute myocardial infarction (I21–I23). The baseline year was 2014, and a 5-year follow-up period was used to predict newly diagnosed CCVDs. To ensure the exclusion of patients with preexisting CCVD diagnoses, the washout period was defined as 2010–2014 (Fig. 1).

Sample selection

The features used in the prediction models included health checkup variables that were conducted biennially. Initially, 418,208 participants with health checkup data available in either 2013 or 2014 and without a CCVD diagnosis between 2010 and 2014 were extracted. Participants with missing major examination variables or eligibility conditions were excluded, resulting in a total study population of 410,859 participants. Two groups were created to develop CCVD prediction models for individuals eligible for both private insurance types. Based on the conditions and operational definitions outlined in Tables 1, 126 and 413 samples met the criteria of the standardized insurance application form (henceforth, "standard group") and 268,912 samples fulfilled the conditions of the simplified enrollment form (henceforth, "simplified group") through restoration sampling. The "standard" group includes individuals who answered "No" to both Question A and Question B, while the "simplified" group includes those who answered "No" to both Question C and Question D. The standard questions are stricter regarding hospitalization, surgery, and prescriptions compared to the simplified questions. Therefore, the simplified group encompasses the standard group. The 15,534 individuals excluded from the

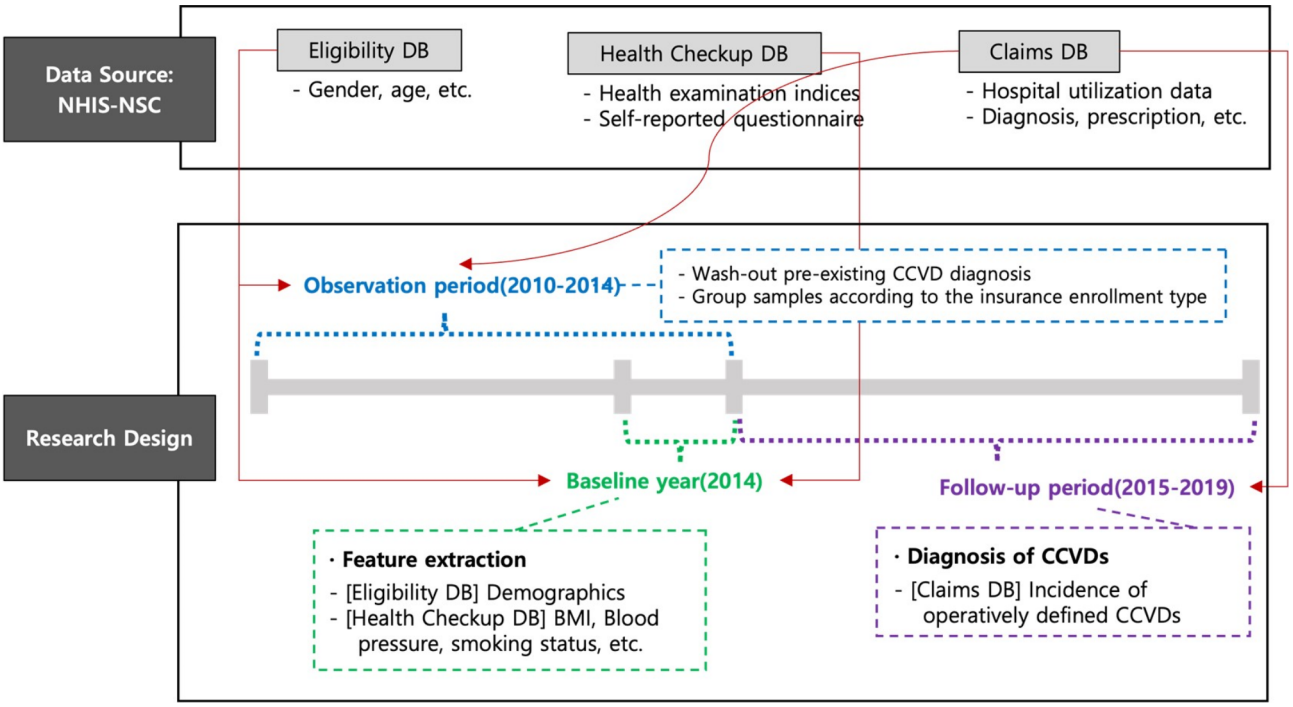

Fig. 1. Data structure and research design.

Screening form	Question
Standard	A. Recent 5-Year medical procedures history
	1) Hospitalization
	2) Operation
	3) Continued treatment for 7 days or more
	B. Recent 5-Year medical procedures for 10 diseases
	1) Diagnosis of confirmed disease
	2) Treatment
	3) Hospitalization
	4) Operation
	5) Medication
Simplified	C. Recent 2-year hospitalization/operation due to disease or injury accident
	D. Recent 5-year diagnosis/hospitalization/operation due to cancer

Table 1. The standardized and simplified private health insurance enrollment form. 10 diseases include cancer, leukemia, hypertension, angina, myocardial infarction, heart valve disease, cirrhosis, stroke, diabetes, AIDS/HIV. The questions above are expected to be answered with 'yes' or 'no'.

sample are those who answered "Yes" to any of the simplified questions. A detailed discussion of screening for simplified and standard private insurance enrollees can be found in the literature[12].

The items in this study were selected for two purposes: to mirror questionnaire responses typically provided by potential insurees to insurers before commencing insurance contracts and to be operationally defined within the dataset. Standard screening form questions typically mandate a longer period, often up to 5 years, during which individuals must have experienced no medical issues, including hospitalizations, surgeries, or diagnosed diseases. By contrast, simplified screening forms designed for individuals with substandard health conditions require shorter intervals without hospitalization or surgery. Disease diagnosis is limited to severe conditions, such as cancer.

Table 2 summarizes the descriptive statistics of the three samples; the incidence rate of the outcomes increased in the order of the standard group, the simplified enrollment group, and those without any screening process (referred to as the 'all' group). In particular, the standard group exhibited an incidence rate of approximately one-third of the overall incidence. The distribution of health checkup indicators and age followed a similar pattern. The standard group samples were characterized by younger age and exhibited lower levels of fasting

	All	Standard	Simplified		All	Standard	Simplified
I. Outcome				II. Features (categorical)			
Cerebrovascular	34,391 (8.37%)	3,200 (2.53%)	17,411 (6.47%)	Gender			
Stroke	21,043 (5.12%)	1,593 (1.26%)	10,418 (3.87%)	Male	208,475 (50.74%)	67,343 (53.27%)	138,625 (51.55%)
Cerebral hemorrhage	2,697 (0.66%)	266 (0.21%)	1,332 (0.50%)	Dipstick test			
Ischemic heart disease	32,896 (8.01%)	2,898 (2.29%)	16,603 (6.17%)	Weak positive	10,746 (2.62%)	3,039 (2.40%)	6,966 (2.59%)
Acute myocardial infarction	3,075 (0.75%)	252 (0.20%)	1,431 (0.53%)	Positive (+1)	6,093 (1.48%)	1,259 (1.00%)	3,740 (1.39%)
II. Features (continuous)				Positive (+2)	2,390 (0.58%)	378 (0.30%)	1,336 (0.50%)
Age (years)	50.18 (14.33)	43.16 (12.11)	48.50 (14.08)	Positive (+3)	668 (0.16%)	90 (0.07%)	324 (0.12%)
Waist Circumference (cm)	80.53 (9.45)	78.84 (9.42)	80.26 (9.49)	Positive (+4)	151 (0.04%)	17 (0.01%)	75 (0.03%)
BMI (kg/m^2)	23.84 (3.35)	23.41 (3.33)	23.81 (3.37)	Family history			
SBP (mmHg)	121.91 (14.71)	118.99 (13.33)	121.65 (14.60)	Stroke	25,467 (6.20%)	6,270 (4.96%)	16,136 (6.00%)
DBP (mmHg)	75.84 (9.90)	74.64 (9.48)	75.82 (9.90)	Heart disease	15,331 (3.73%)	4,551 (3.60%)	9,991 (3.72%)
Hemoglobin (g/dL)	14.02 (1.63)	14.18 (1.66)	14.09 (1.63)	Hypertension	56,527 (13.76%)	14,716 (11.64%)	37,189 (13.83%)
Fasting blood serum (mg/dL)	99.16 (24.01)	93.95 (15.10)	98.42 (22.95)	Diabetes	42,806 (10.42%)	12,513 (9.90%)	28,160 (10.47%)
Total Cholesterol (mg/dL)	194.83 (39.18)	195.75 (37.39)	195.16 (38.99)	Cancer/etc	56,411 (13.73%)	16,811 (13.30%)	35,907 (13.35%)
Triglyceride (mg/dL)	129.71 (94.64)	122.72 (92.77)	129.18 (95.18)	Smoking status			
HDL-Cholesterol (mg/dL)	55.43 (15.28)	56.76 (15.35)	55.61 (15.06)	Quit smoking	64,262 (15.64%)	17,248 (13.64%)	40,731 (15.15%)
LDL-Cholesterol (mg/dL)	114.38 (41.91)	115.62 (43.42)	114.74 (42.96)	< 20 cigarettes/day	54,239 (13.20%)	21,103 (16.69%)	38,028 (14.14%)
Serum creatinine (mg/dL)	0.88 (0.41)	0.87 (0.35)	0.88 (0.38)	≥ 20 cigarettes/day	37,179 (9.05%)	11,572 (9.15%)	24,177 (8.99%)
SGOT (U/L)	25.47 (17.85)	23.86 (13.93)	24.97 (15.59)	Drinking status			
SGPT (U/L)	24.95 (24.43)	23.86 (22.46)	24.69 (22.22)	Drinker (≥ 2 times/week)	55,619 (13.54%)	16,204 (12.82%)	36,089 (13.42%)
GGT (U/L)	35.98 (49.49)	32.65 (39.95)	35.11 (45.74)	Exercise			
eGFR ($mL/min/1.73m^2$)	90.15 (23.57)	93.19 (22.09)	90.76 (23.03)	More than once a week	378,004 (92.00%)	113,879 (90.08%)	246,036 (91.49%)
Sample size (N)	410,859	126,413	268,912	Sample size (N)	410,859	126,413	268,912

Table 2. Summary statistics. Abbreviations: SBP for systolic blood pressure, DBP for diastolic blood pressure, HDL for high-density lipoprotein, LDL for low-density lipoprotein, SGOT for serum glutamic-oxaloacetic transaminase, SGPT for serum glutamic pyruvic transaminase, GGT for gamma-glutamyltranspeptidase, eGFR for estimated glomerular filtration rate. Continuous variables are reported as mean(sd), while category variable statistics are presented as N(%). Gender, and dipstick test variables are categorized as female/male and negative/weak positive/positive(+1)/positive(+2)/positive(+3)/positive(+4), respectively. Each items of the family history questionnaire is expected to be answered by no/yes. Smoking status, alcohol consumption and exercise is categorized by non-smoker/currently quit smoking/smokes less than 20 cigarettes per day/smokes more or equal to 20 cigarettes per day, drinks once or less per week/drinks twice or more per week, none/moderate or vigorous exercise more than once a week, respectively. The number and the proportion of the first category is omitted in the table.

blood serum triglycerides, systolic blood pressure (SBP), and various other health checkup indicators or family history records than the other groups.

Variables

Predictor variables were selected as follows: sex and age information were extracted from the eligibility database. Fifteen variables were obtained from the checkup database, including BMI, waist circumference, SBP, diastolic blood pressure, hemoglobin, fasting blood serum, total cholesterol, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, triglyceride, serum glutamic oxaloacetic transaminase, serum glutamic pyruvic transaminase, gamma-glutamyl transferase, urine dipstick test, serum creatinine, and estimated glomerular filtration rate.

Variables with specified thresholds were further processed to eliminate any variations in risk within the criteria defining the normal range for each index (Supplementary Table 1). Variables within the normal range were assigned a value of 0, whereas those beyond the normal range were adjusted to represent the absolute difference from the threshold value. This approach was adopted to estimate the incremental changes in risk beyond the normal range and capture the transition from no risk variation within the normal range to a potentially hazardous range. When the normal range provided was one-sided, a single variable indicating a deviation from the threshold value of the normal range was generated. For variables with a two-sided range, two variables representing the deviations from the minimum and maximum boundaries were generated. For example, if the normal range for BMI is 18.5–22.9, two variables—LOW_BMI and HIGH_BMI—were created to represent the distances from 18.5 to 22.9, respectively.

Statistical analysis

To predict the 5-year cumulative incidence of CCVDs, we utilized logistic regression and penalized logistic regression models, which offer the advantage of interpretability compared with black-box models. While complex models may improve predictive power, they often obscure the relationships between risk factors and outcomes. Logistic regression, widely used for binary outcomes, provides clear insights into variable importance, making it well-suited for clinical applications, without any severe loss of predictive ability.

To predict the 5-year cumulative incidence of CCVDs, we utilized logistic regression and penalized logistic regression models, which offer the advantage of interpretability compared with blackbox models. Logistic regression is widely used to predict binary outcomes using multiple variables. The logistic regression model is represented by the following equation:

$$y_i = \log\left(\frac{p_i}{1-p_i}\right) = X_i\beta + \epsilon_i$$

By examining the signs and magnitudes of the coefficient vector (β) in the model, the marginal effects of each risk factor can be calculated, enabling a straightforward interpretation of the impact that each predictor has on the likelihood of developing CCVDs. Penalized logistic regression further enhances the prediction performance by incorporating l_1 or l_2 penalty terms into the objective function, reducing the size of the coefficient vector, and addressing the issue of overfitting. A nonnegativity constraint was also imposed on the coefficient vector, except for sex, as the features were preprocessed to indicate the absolute distance from the normal range of each health checkup variable, as previously discussed. Therefore, the objective function is given by:

$$\widehat{\beta} = argmin_\beta \left\{\frac{1}{2}\sum\nolimits_i^N (y_i - X_i\beta)^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2\right\} \; s.t.\; \beta \geq 0$$

In the LASSO model, the l_1 penalty (λ_1) shrinks some coefficients to zero, allowing for the identification of the most influential predictors. This feature facilitates variable selection and the identification of key risk factors associated with CCVDs. Conversely, the ridge model, with its l_2 penalty (λ_2), reduces the size of the coefficients without reducing them to zero. These regularization techniques help reduce data dependency and improve the generalizability of the model. Both models provide interpretable results while effectively addressing the issues of multicollinearity and overfitting[13–15]. For model training and validation, the study population was divided into an 80% training set and a 20% test set. The training set was further divided into 10 folds for hyperparameter tuning using 10-fold cross-validation (Fig. 2).

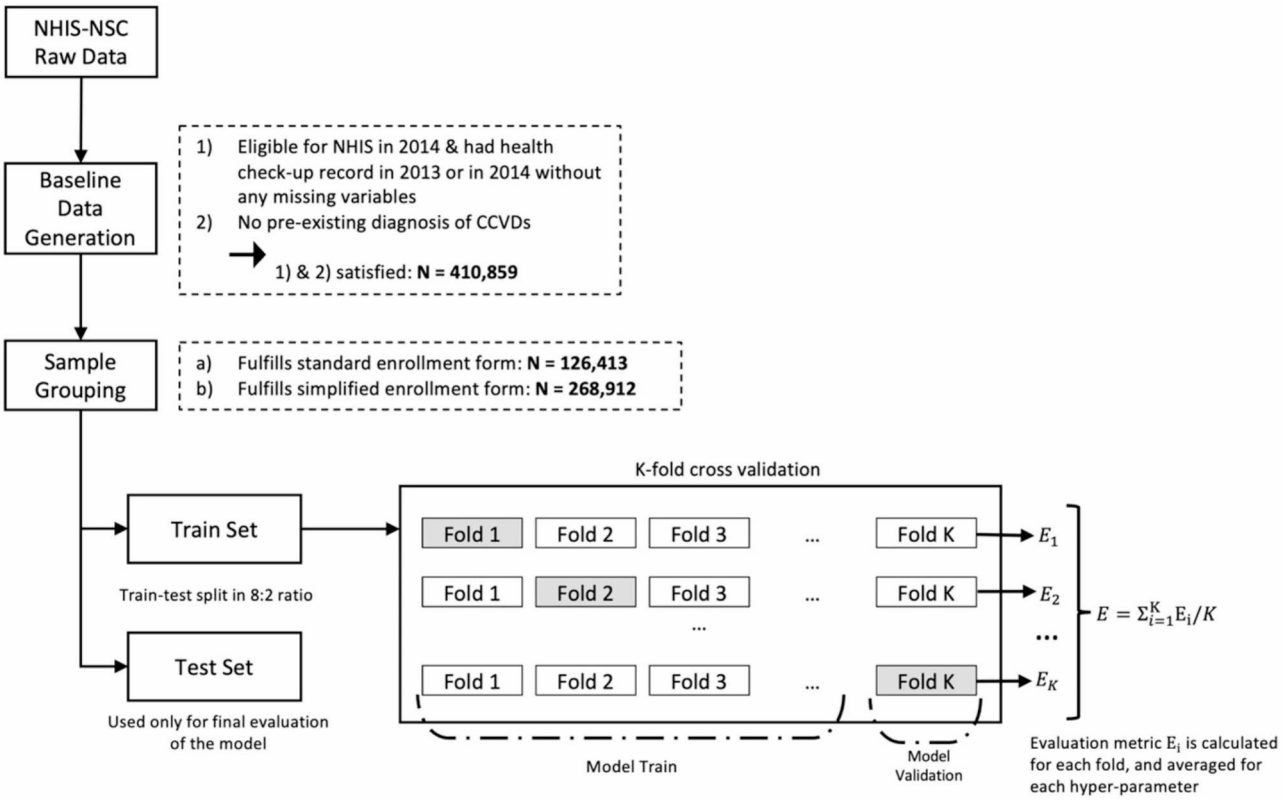

Fig. 2. Sample selection, and data training process.

Evaluation metrics, such as the area under the receiver operating characteristic curve (AUROC), accuracy ([TN + TP]/[TN + FN + FP + TP]), recall (TP/[TP + FN]), and specificity (TN/[TN + FP]), were employed to compare the predictive powers between samples and models, where TN, TP, FN, and FP are the number of true negatives, true positives, false negatives, and false positives, respectively. Given the threshold dependency of these metrics and the imbalanced distribution of the outcome variables, a cutoff value was selected to optimize both recall and specificity. Specifically, the threshold was selected from the upper-left part of the receiver-operator curve, where the distance from the diagonal line was maximized[16].

Results

Predictive performance

We use two samples in this analysis: the standard group and the simplified group. The standard group underwent a stricter screening process, resulting in a lower CCVD incidence rate, younger age distribution, and generally healthier checkup indices. In contrast, the simplified group exhibited a higher incidence rate and a less favorable health profile. These differences in initial risk factors, influenced by the screening process, are expected to impact the predictive performance of our models. Accordingly, we fitted and trained the models separately for each group. The training process for the models in each group was initiated via feature selection using the LASSO algorithm. Subsequently, logistic and ridge regressions were conducted separately to fit the selected variables. For comparison, logistic regression without penalty was also performed. Table 3 provides a comprehensive comparison of the predictive performances of these models. Within the table, "Logit" refers to a model trained with logistic regression without any penalty; "LASSO" refers to a model trained solely using the LASSO algorithm; and "LASSO & Logit" and "LASSO & Ridge" refer to models that employ logistic and ridge regression on the predictors selected by the LASSO algorithm, respectively. For a comprehensive evaluation of the models, AUROC, accuracy, recall, and specificity were used as evaluation metrics. After assessing the overall predictive ability, the optimal threshold to calculate accuracy, recall, and specificity was selected using the AUROC and its confidence interval[17], regardless of the cutoff value of the models. Finally, the geometric mean of recall and specificity was calculated to evaluate the overall classification ability of the model.

When comparing the evaluation metrics of the four models, specifically focusing on the AUROC, both logistic regression without any penalty and nonnegativity-constrained logistic regression with variables selected using the LASSO algorithm exhibited superior predictive performance. However, it is worth noting that the differences in these performance measures were statistically insignificant. This trend remained consistent regarding the geometric mean. Except for cerebral hemorrhage in the standard group and ischemic heart disease in the simplified group, these two models generally outperformed the other models.

		Standard						Simplified					
Outcome	Model	AUROC (CI)	Cut-off	Acc	Rec	Spec	G. Mean	AUROC (CI)	Cut-off	Acc	Rec	Spec	G. Mean
CBVD	Logit	0.753 (0.734–0.772)	0.026	0.696	0.702	0.696	0.699	0.791 (0.784–0.799)	0.050	0.646	0.810	0.634	0.717
	Lasso	0.748 (0.729–0.767)	0.026	0.611	0.777	0.607	0.687	0.790 (0.783–0.797)	0.056	0.667	0.780	0.660	0.717
	Lasso & Logit	0.753 (0.734–0.772)	0.026	0.694	0.705	0.694	0.699	0.791 (0.784–0.798)	0.053	0.662	0.790	0.653	0.718
	Lasso & Ridge	0.746 (0.727–0.765)	0.026	0.688	0.696	0.688	0.692	0.790 (0.782–0.797)	0.069	0.699	0.745	0.696	0.720
Stroke	Logit	0.783 (0.757–0.809)	0.010	0.623	0.810	0.621	0.709	0.812 (0.803–0.820)	0.043	0.755	0.727	0.756	0.742
	Lasso	0.775 (0.748–0.801)	0.011	0.629	0.796	0.627	0.706	0.809 (0.800-0.817)	0.039	0.719	0.758	0.718	0.738
	Lasso & Logit	0.784 (0.757–0.810)	0.011	0.663	0.777	0.661	0.717	0.812 (0.803–0.820)	0.036	0.712	0.776	0.709	0.742
	Lasso & Ridge	0.782 (0.755–0.808)	0.014	0.684	0.752	0.683	0.717	0.809 (0.800-0.817)	0.045	0.741	0.738	0.741	0.739
CH	Logit	0.726 (0.639–0.814)	0.003	0.767	0.625	0.767	0.692	0.774 (0.748-0.800)	0.005	0.662	0.773	0.661	0.715
	Lasso	0.710 (0.617–0.804)	0.003	0.736	0.650	0.736	0.692	0.764 (0.737–0.790)	0.005	0.622	0.788	0.621	0.700
	Lasso & Logit	0.725 (0.635–0.814)	0.002	0.697	0.675	0.697	0.686	0.773 (0.747–0.799)	0.005	0.707	0.723	0.707	0.715
	Lasso & Ridge	0.705 (0.610–0.801)	0.002	0.854	0.500	0.854	0.653	0.769 (0.743–0.795)	0.005	0.667	0.759	0.666	0.711
IHD	Logit	0.736 (0.716–0.755)	0.024	0.673	0.690	0.673	0.681	0.773 (0.766–0.781)	0.054	0.653	0.782	0.645	0.710
	Lasso	0.729 (0.709–0.749)	0.021	0.589	0.765	0.585	0.669	0.772 (0.764–0.779)	0.054	0.634	0.804	0.623	0.708
	Lasso & Logit	0.738 (0.718–0.757)	0.023	0.657	0.715	0.656	0.685	0.773 (0.765–0.780)	0.052	0.639	0.795	0.629	0.707
	Lasso & Ridge	0.733 (0.714–0.752)	0.023	0.587	0.783	0.582	0.675	0.772 (0.764–0.779)	0.059	0.637	0.796	0.627	0.706
AMI	Logit	0.812 (0.761–0.863)	0.002	0.708	0.815	0.708	0.760	0.828 (0.808–0.848)	0.005	0.708	0.815	0.708	0.759
	Lasso	0.786 (0.734–0.838)	0.002	0.598	0.852	0.597	0.713	0.822 (0.802–0.843)	0.006	0.725	0.790	0.724	0.756
	Lasso & Logit	0.807 (0.757–0.857)	0.001	0.662	0.870	0.662	0.759	0.826 (0.806–0.846)	0.005	0.682	0.847	0.681	0.759
	Lasso & Ridge	0.792 (0.741–0.843)	0.002	0.724	0.778	0.724	0.751	0.821 (0.801–0.841)	0.006	0.679	0.836	0.679	0.753

Table 3. Evaluation metrics for all models. Abbreviation: AUROC stands for area under receiver-operator curve, Acc for accuracy, Rec for recall, Spec for specificity, G. Mean for geometric mean of recall and specificity. CBVD refers to cerebrovascular disease, CH to cerebral hemorrhage, IHD to ischemic heart disease, and AMI to accute myocardial infarction. The 95% confidence interval for AUROC was determined using the DeLong's test. The optimal cut-off point is selected from the receiver-operator curve to maximize the distance from the diagonal line following Youden's method.

The AUROC of the simplified group—with a larger number of observations and incident cases—generally outperformed that of the standard group. This trend was also reflected in the geometric mean of recall and specificity, indicating that models of simplified samples were more balanced regarding the tradeoff between accuracy in positive and negative cases.

Regression coefficients

The coefficients that exhibited the highest overall AUROC values, estimated by the logistic regression models using variables selected by the LASSO algorithm, are shown in Table 4. Although logistic regression without constraints also demonstrated strong predictive performance, we encountered an issue with negative coefficients in this model. Because all variables were treated as risk factors, the presence of negative coefficients contradicted our expectations. Consequently, we focused our discussion solely on the regression results obtained from the logistic regression with variable selection.

Variables excluded through the LASSO algorithm for certain outcomes are denoted by a dash (-), whereas variables excluded for all outcomes are not included in the table. Several notable findings emerged from this study. Overall, models with heart disease outcomes (IHD and AMI) and those fitted to the standard group experienced more frequent variable exclusions. Furthermore, when comparing the coefficient values across different diseases, apart from cerebral hemorrhage, the simplified group consistently exhibited higher coefficients for all diseases. This implies that within the simplified enrollment group, the estimated incidence rate experiences a more significant increase with each unit increase in risk factors, such as screening indicators (e.g., blood pressure or age). This suggests that the same increase in these risk factors results in a greater escalation of risk within the simplified enrollment group, which already exhibits a higher prevalence of existing risk factors. These findings highlight the interplay between risk factors, disease incidence, and choice of insurance enrollment type. Specifically, individuals within the simplified enrollment group, who often possess a greater burden of risk factors, tend to experience a more pronounced increase in their estimated risk of various diseases.

Variable importance

Table 5 shows the variable importance of the logistic regression for the selected variables. The variable importance for each feature was calculated by conducting a logistic regression while normalizing the variables. Age has consistently emerged as the most important variable across all models. This finding is reasonable, considering that age is a factor that depreciates other health assets that are not measured in health checkups and self-reported questionnaires. Considering this, the importance of variables other than age, as shown in the table, is presented as relative importance; the importance of age was set at 100.

A notable finding demonstrated in the table is the elevated importance of variables related to smoking status and family history of stroke. Additionally, sex, BMI, and SBP showed high levels of importance. When comparing the two groups, the rankings of variable importance exhibited similar patterns for all variables, except

	Standard					Simplified				
	CBVD	Stroke	CH	IHD	AMI	CBVD	Stroke	CH	IHD	AMI
Gender	-0.216	0.018	0.010	0.212	0.549	-0.046	0.190	0.139	0.234	0.692
Age	0.072	0.077	0.055	0.059	0.076	0.080	0.085	0.060	0.068	0.073
High BMI	0.019	0.036	0.042	0.042	-	0.023	0.023	0.010	0.064	0.047
High SBP	0.009	0.014	0.023	0.005	0.026	0.003	0.007	0.008	0.001	-
High DBP	0.005	0.007	0.031	0.013	0.006	0.008	0.010	0.022	0.006	0.013
Low HGB	-	-	0.381	-	-	-	0.023	0.172	-	-
High FBS	0.001	0.002	-	0.001	0.007	0.002	0.003	0.001	0.002	0.005
Low HDL	0.004	0.003	-	0.009	0.026	0.005	0.005	0.003	0.011	0.023
High SGOT	0.002	0.002	0.002	-	-	-	0.000	0.002	-	-
High SGPT	-	0.000	-	-	-	0.000	0.001	-	-	-
High GGT	0.001	0.002	0.003	-	-	0.001	0.001	0.002	-	-
High CRTN	-	-	-	-	-	0.032	0.031	0.064	-	-
Low eGFR	0.022	0.019	0.037	-	-	0.006	0.007	0.010	0.013	0.024
Dipstick test	-	-	0.113	0.070	-	0.088	0.105	0.149	0.084	0.024
Smoker	0.055	0.094	0.113	0.044	0.507	-	0.015	0.006	0.059	0.244
High risk drinker	0.089	0.180	0.294	-	-	0.028	0.064	0.135	-	-
Family history: stroke	0.405	0.391	0.591	0.227	-	0.445	0.506	0.354	0.230	0.170
Family history: heart disease	0.152	0.361	-	0.334	0.610	0.059	0.083	-	0.606	0.622
Family history: hypertension	0.059	0.200	-	0.172	-	0.192	0.215	0.200	0.247	0.153
Constant	-7.331	-8.629	-9.531	-7.092	-11.936	-7.386	-8.536	-8.629	-7.093	-10.689

Table 4. Logistic regression coefficients for variables selected by Lasso algorithm in standard and simplified groups. Abbreviations are in accordance with those found in Tables 2 and 4. Variables not selected by the Lasso algorithm in any case are excluded from this table.

	Standard					Simplified				
	CBVD	Stroke	CH	IHD	AMI	CBVD	Stroke	CH	IHD	AMI
Gender	8.708	0.281	0.000	10.192	17.751	1.400	7.705	6.841	10.572	28.746
High BMI	3.731	8.019	14.937	12.057	-	4.962	5.483	2.644	18.852	11.978
High SBP	6.195	10.644	26.774	2.975	20.512	2.152	6.036	9.433	0.000	-
High DBP	1.015	2.454	21.899	5.039	0.775	2.499	4.203	13.919	2.111	7.264
Low HGB	-	-	41.153	-	-	-	0.939	11.989	-	-
High FBS	0.837	2.544	-	1.253	15.123	5.247	9.177	1.893	5.533	21.070
Low HDL	2.688	2.249	-	8.468	22.517	3.534	4.536	2.969	10.814	21.826
High SGOT	2.864	1.387	5.330	-	-	-	0.000	6.668	-	-
High SGPT	-	0.000	-	-	-	0.000	0.911	-	-	-
High GGT	4.679	7.995	29.530	-	-	2.629	2.582	14.093	-	-
High CRTN	-	-	-	-	-	0.360	1.066	3.816	-	-
Low eGFR	5.024	4.367	19.468	-	-	1.437	2.353	4.853	4.706	11.789
Dipstick test	-	-	6.427	1.971	-	3.489	5.158	10.129	3.718	0.000
Smoker	4.097	7.874	13.751	3.737	51.699	-	1.047	-	4.809	24.014
High risk drinker	2.372	6.108	15.497	-	-	0.101	1.993	5.405	-	-
Family history: stroke	12.008	11.178	25.602	6.720	-	12.825	15.183	13.094	6.704	3.833
Family history: heart disease	2.599	7.860	-	8.595	13.239	0.326	1.440	-	15.567	14.618
Family history: hypertension	1.087	6.236	-	5.872	-	6.310	7.567	7.103	9.748	4.399

Table 5. Variable importance of logistic regression for variables selected by Lasso algorithm in standard and simplified groups. Abbreviations are in accordance with those found in Tables 2 and 4. Variables not selected by the Lasso algorithm in any case are excluded from this table.

cerebrovascular disease. Notably, for ischemic heart disease, the variable importance scores were remarkably similar between the two groups.

Discussion

This study employed various linear models to predict the incidence of CCVDs in two distinct samples. The logistic regression model with penalty terms demonstrated a superior fit regarding predictive performance compared to other penalized models, albeit without statistical significance. Furthermore, the model with the selected variables and penalty exhibited superior explainability compared to the model without any penalty. Notably, the overall model performance improved as the incidence rate of the samples and the prevalence of the target diseases increased.

These findings underscore the significance of customizing risk assessments to accommodate nuanced variations across individuals with different initial health statuses and diseases with varying risk factors. Such tailored approaches hold promise for enhancing disease prediction and developing insurance products. Moreover, they emphasize the need for further research and the refinement of predictive modeling techniques tailored to specific health conditions within various demographic groups.

Furthermore, by analyzing regression coefficients that are not excluded, it is essential to secure smoking habits or family history information related to the same area in advance to predict the risk factors for cardiovascular and cerebrovascular diseases. This is because, in almost all models, the family history of the risk-secured area appears as a variable with great significance. Several clinical studies have shown that family history is a major risk factor for cardiovascular disease[18,19]. Previous studies related to the predictive model in Korea did not consider family history[7,10]; therefore, accurate comparisons could not be made. However, in overseas studies, family history was a significant variable in the cardiovascular disease prediction model using machine learning[20].

In situations where the prevalence of cardiovascular disease and medical expenses continue to increase, it is essential to secure a quantitative and objective basis. Cardiovascular diseases account for approximately 17% of Korea's total medical expenses, and the burden of medical expenses due to cardiovascular diseases is significant[4]. In the United States, the estimated total cost of cardiovascular disease and related costs as of 2010 was $315.4 billion[21], while in Russia, cardiovascular disease-related costs accounted for 0.19% of the gross domestic product as of 2009[22].

These findings have significant implications for public health policy and insurance risk management. First, the predictive models developed in this study can support personalized health management and policy interventions by assessing CCVD risk in advance. They enable early identification of high-risk individuals based on health screening data, facilitating timely preventive measures such as lifestyle modifications and medical treatments. Raising awareness of personal health risks can also encourage proactive health behaviors, ultimately improving public health and reducing healthcare costs.Second, these models can enhance insurers' risk assessment and product development. By integrating predictive modeling into underwriting, insurers can quantify health risks, adjust premiums accordingly, and offer incentives for health management. For instance, premium discounts for high-risk individuals engaging in health programs can reduce insurer risk while promoting policyholders' health, leading to more precise risk management and tailored insurance products.

This study has some limitations. First, as CCVDs were identified solely based on the International Classification for Diseases, Version 10 (ICD-10) codes, not all individuals with cardiovascular diseases were identified. Second, in the analysis group definition stage, the exclusion of low and high age groups was significant. This is related to the characteristics of patients undergoing health examinations, and there are reasons for their weakness regarding the purpose of prevention and access to examination reservations. This caused bias in this study, and it will thus be necessary to supplement our findings through customized DBs—such as elderly cohort DBs—for analyses that control for age in future studies.

Third, the data used in this study exhibit class imbalance between CCVD occurrence and non-occurrence cases. This imbalance may affect the model's predictive performance, particularly recall, and performance improvement could be expected through techniques such as oversampling. However, in this study, oversampling was not applied to maintain the interpretability of the model and the consistency of variable selection. This remains a limitation that should be considered when interpreting the results.

Finally, although this study accounted for various confounders measured in the data, the possibility of residual confounding cannot be completely ruled out. Despite these limitations, this study is noteworthy because it compares the predictions of various models using the latest data.

Conclusion

This study utilized NHIS sample cohort data to predict CCVDs incidence and compare model performance across different health conditions. The logistic regression model with LASSO-selected variables showed the best predictive performance, particularly in the simplified group with higher incidence rates and risk levels. Age, smoking status, family history of stroke, sex, BMI, and systolic blood pressure were identified as key risk factors, emphasizing the need for personalized risk assessment and diverse predictive models for effective CCVDs prevention and management.

This study contributes to CCVDs prevention and management strategies through several strengths. First, using NHIS data, the study developed a robust predictive model representative of the Korean population, enhancing generalizability. Second, by classifying samples into standard and simplified groups, key risk factors were identified based on health status. Third, the application of penalized linear models, balancing interpretability and predictive performance, enhances practical utility. Collectively, these findings provide scientific evidence to support effective prevention and management strategies for CCVDs.

Data availability

Raw data were obtained from the Korean National Health Insurance System. The data supporting the findings of this study are available from the corresponding author upon request.

Received: 30 July 2024; Accepted: 17 March 2025
Published online: 02 July 2025

References

1. Read, S. H. & Wild, S. H. Prevention of premature cardiovascular death worldwide. *Lancet (London England)*. **395** (10226), 758–760 (2020).
2. Safiri, S. et al. Global, regional and National burden of osteoarthritis 1990–2017: A systematic analysis of the global burden of disease study 2017. *Ann. Rheum. Dis.* **79** (6), 819–828 (2020).
3. Kyu, H. H. et al. Global, regional, and National disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: A systematic analysis for the global burden of disease study 2017. *Lancet* **392** (10159), 1859–1922 (2018).
4. Korean Statistical Information Service. Annual Report on the Cause of Death Statistics, Daejeon, Korea: Statistics Korea, (2022).
5. Artinian, N. T. et al. Interventions to promote physical activity and dietary lifestyle changes for cardiovascular risk factor reduction in adults: A scientific statement from the American heart association. *Circulation* **122** (4), 406–441 (2010).
6. Lee, S., Cho, E., Jeon, S. & Hong, S. C. Predicting the risk of major chronic diseases and its application: Using NHIS big data. *Korean J. Insurance*. **133**, 23–48 (2023).
7. Lee, S. J. et al. Deep learning improves prediction of cardiovascular disease-related mortality and admission in patients with hypertension: Analysis of the Korean National health information database. *J. Clin. Med.* **11** (22), 6677 (2022).
8. Dinh, A., Miertschin, S., Young, A. & Mohanty, S. D. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med. Inf. Decis. Mak.* **19** (1), 1–15 (2019).
9. Goldstein, B. A., Navar, A. M. & Carter, R. E. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *Eur. Heart J.* **38** (23), 1805–1814 (2017).
10. Joo, G., Song, Y., Im, H. & Park, J. Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (Nationwide cohort data in Korea). *IEEE Access*. **8**, 157643–157653 (2020).
11. Lee, J., Lee, J. S., Park, S. H., Shin, S. A. & Kim, K. Cohort profile: The National health insurance service–national sample cohort (NHIS-NSC), South Korea. *Int. J. Epidemiol.* **46** (2), e15–e15 (2017).
12. Hwang, J. Y., Lee, S. Y. & Joo, S. H. Optimal underwriting questionnaire calculation of simplified issue product by coverage using National health insurance data. *Korean J. Insurance*. **124**, 1–34 (2020).
13. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *Annals Stat.* **32** (2), 407–499 (2004).
14. Gaines, B. R., Kim, J. & Zhou, H. Algorithms for fitting the constrained Lasso. *J. Comput. Graphical Stat.* **27** (4), 861–871 (2018).
15. Hu, Z., Follmann, D. A. & Miura, K. Vaccine design via nonnegative lasso-based variable selection. *Stat. Med.* **34** (10), 1791–1798 (2015).
16. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3** (1), 32–35 (1950).
17. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 837–845. (1988).
18. Psaltopoulou, T. et al. Socioeconomic status and risk factors for cardiovascular disease: Impact of dietary mediators. *Hellenic J. Cardiol.* **58** (1), 32–42 (2017).
19. Kinoshita, M. et al. Japan atherosclerosis society (JAS) guidelines for prevention of atherosclerotic cardiovascular diseases 2017. *J. Atheroscler. Thromb.* **25** (9), 846–984 (2018).

20. Abdalrada, A. S., Abawajy, J., Al-Quraishi, T. & Islam, S. M. S. Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: A retrospective cohort study. *J. Diabetes Metab. Disord.* **21** (1), 251–261 (2022).
21. Korsnes, J. S., Davis, K. L., Ariely, R., Bell, C. F. & Mitra, D. Health care resource utilization and costs associated with nonfatal major adverse cardiovascular events. *J. Manag. Care Specialty Pharm.* **21** (6), 443–450 (2015).
22. Kontsevaya, A., Kalinina, A. & Oganov, R. Economic burden of cardiovascular diseases in the Russian federation. *Value Health Reg. Issues.* **2** (2), 199–204 (2013).

Acknowledgements
None.

Author contributions
Literature search: Sehyun Kim, Mingee Choi; writing—original draft: Mingee Choi, Sangyon Lee; writing—review and editing: Beomsang Ryu, Mingee Choi, Sangyon Lee; data analysis: Beomsang Ryu; validation: Beomsang Ryu; study design: Jaeyong Shin, Sok Chul Hong; methodology: Sok Chul Hong; data interpretation: Jaeyong Shin, & Sok Chul Hong; supervision: Jaeyong Shin, & Sok Chul Hong; project administration: Jaeyong Shin, & Sok Chul Hong.

Funding
This research was supported by the Academic Activity Support Program for Faculty at Seoul National University.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval and consent to participate
This study used secondary data, in which all personal information was anonymized and de-identified. Hence, obtaining patient consent was waived off. Ethical approval was waived by the Institutional Review Board of Seoul National University (IRB No. P01-202206-01-031).

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-94888-0.

Correspondence and requests for materials should be addressed to J.S. or S.C.H.