Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: Yunjie Chai

Date: 07/28/2025

Approval Sheet

Large The Joint Effects of Diabetes and Depression on Stroke in the Atherosclerosis Risk in Communities (ARIC) Study

By

Yunjie Chai
MPH
Epidemiology

Committee Chair: Alvaro Alonso

Abstract Cover Page

Large The Joint Effects of Diabetes and Depression on Stroke in the Atherosclerosis Risk in Communities (ARIC) Study

By

Yunjie Chai
Bachelor of Science
UC Davis
2017

Thesis Committee Chair: Professor Alvaro Alonso

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2025

Abstract

Large The Joint Effects of Diabetes and Depression on Stroke in the Atherosclerosis Risk in Communities (ARIC) Study

By Yunjie Chai

1 Abstract

1.1 Background and Objectives

Although diabetes and depression are well recognized risk factors for stroke, their joint effects on stroke risk remain underexplored. This study used data from the Atherosclerosis Risk in Communities (ARIC) Study to quantify combined associations of prestroke depression and diabetes with stroke risk, and to investigate the predictive potential of depression-diabetes multimorbidity on stroke risk using logistic regression and machine learning models.

1.2 Methods

We analyzed 5,459 ARIC participants who attended Visit 5 (2011–2013), were free of prior stroke, and had complete data on diabetes and depressive symptoms (CES-D ≥9). Participants were classified into four exposure groups: no diabetes or depression, diabetes only, depression only, and both diabetes and depression. Incident stroke events were ascertained through 2020 via adjudicated hospitalization records. Cox proportional hazards models estimated hazard ratios and 95% confidence intervals across three adjustment levels: unadjusted, adjusted for age, sex, and race, and fully adjusted. We then applied XGBoost classifier and stepwise logistic regression (AIC-based). Model discrimination was assessed by AUC.

1.3 Result

Over a median 7.8-year follow-up, 233 incident strokes occurred. In the fully adjusted Cox model, individuals with both depression and diabetes had the highest risk of stroke (HR 1.95; 95% CI 1.05–3.60), compared to reference. Diabetes only (HR 1.22; 95% CI 0.91–1.63) and depression only (HR 1.12; 95% CI 0.55–2.30) showed weaker, non-significant associations. The XGBoost model achieved moderate discrimination, improving recall for stroke cases from 0.40 to 0.62 through threshold adjustment. The final logistic regression model included age, prevalent CHD, the depression×diabetes interaction (p=0.025), and hypertension, yielding a C-statistic of 0.625 (95%CI 0.59–0.66).

1.4 Conclusion

Prestroke depression-diabetes multimorbidity suggests a synergistic increase in stroke risk beyond individual effects. However, the stroke risk predictive performance of the models was constrained by class imbalance and limited predictor scope. These findings demonstrates the importance of integrated metabolic and mental health management to mitigate stroke risk.

Cover Page

Large The Joint Effects of Diabetes and Depression on Stroke in the Atherosclerosis Risk in Communities (ARIC) Study

By
Yunjie Chai
Bachelor of Science
UC Davis
2017

Thesis Committee Chair: Professor Alvaro Alonso

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2025

The Joint Effects of Diabetes and Depression on Stroke in the Atherosclerosis Risk in Communities (ARIC) Study

Chai Yunjie

July 28, 2025

Contents

1 Abstract

1.1 Background and Objectives

Although diabetes and depression are well recognized risk factors for stroke, their joint effects on stroke risk remain underexplored. This study used data from the Atherosclerosis Risk in Communities (ARIC) Study to quantify combined associations of prestroke depression and diabetes with stroke risk, and to investigate the predictive potential of depression-diabetes multimorbidity on stroke risk using logistic regression and machine learning models.

1.2 Methods

We analyzed 5,459 ARIC participants who attended Visit 5 (2011–2013), were free of prior stroke, and had complete data on diabetes and depressive symptoms (CES-D $\geq$9). Participants were classified into four exposure groups: no diabetes or depression, diabetes only, depression only, and both diabetes and depression. Incident stroke events were ascertained through 2020 via adjudicated hospitalization records. Cox proportional hazards models estimated hazard ratios and 95% confidence intervals across three adjustment levels: unadjusted, adjusted for age, sex, and race, and fully adjusted. We then applied XGBoost classifier and stepwise logistic regression (AIC-based). Model discrimination was assessed by AUC.

1.3 Result

Over a median 7.8-year follow-up, 233 incident strokes occurred. In the fully adjusted Cox model, individuals with both depression and diabetes had the highest risk of stroke (HR 1.95; 95% CI 1.05–3.60), compared to reference. Diabetes only (HR 1.22; 95% CI 0.91–1.63) and depression only (HR 1.12; 95% CI 0.55–2.30) showed weaker, non-significant associations. The XGBoost model achieved moderate discrimination, improving recall for stroke cases from 0.40 to 0.62 through threshold adjustment. The final logistic regression model included age, prevalent CHD, the depression$\times$diabetes interaction (p=0.025), and hypertension, yielding a C-statistic of 0.625 (95%CI 0.59–0.66).

1.4 Conclusion

Prestroke depression-diabetes multimorbidity suggests a synergistic increase in stroke risk beyond individual effects. However, the stroke risk predictive performance of the models was constrained by class imbalance and limited predictor scope. These findings

demonstrates the importance of integrated metabolic and mental health management to mitigate stroke risk.

2 Introduction

Stroke remains one of the leading causes of long-term disability and death worldwide (World Health Organization, 2025). According to FastStats from the CDC, there were 162,639 deaths attributed to cerebrovascular disease in the United States in 2023, making stroke the fourth leading cause of death nationally (Centers for Disease Control and Prevention, 2024). Beyond its human toll, stroke imposes a substantial economic burden. The direct medical costs in the United States for stroke were estimated at $52.8 billion for 2017–2018, including the cost of healthcare services, medications, and missed days of work (Tsao et al., 2022).

Among established vascular risk factors, both diabetes and depression play significant roles. Diabetes mellitus contributes to chronic inflammation, endothelial dysfunction, and metabolic derangement, which accelerate vascular damage (R. Chen et al., 2016). Major depression has been associated with behavioral dysregulation, hypothalamic–pituitary–adrenal (HPA) axis disturbance, and elevated inflammatory cytokines (R. Chen et al., 2016; Emerging Risk Factors Collaboration, 2010). Each condition independently increases the risk of stroke through pathways such as chronic inflammation, HPA axis, and autonomic dysregulation, metabolic dysfunction, and adverse health behaviors (Pan et al., 2011; Sacco et al., 2015). These overlapping mechanisms suggest the potential for synergistic effects on vascular health when both conditions are present.

Evidence from large cohort studies supports the adverse impact of multimorbidity, particularly involving diabetes and depression, on stroke risk and post-stroke outcomes. Ouk et al. (Ouk et al., 2020) found that patients with both conditions had substantially higher risks of institutionalisation and post-stroke dementia, particularly among women. A meta-analysis also indicated that depression is associated with a 1.5-fold increase in all-cause and cardiovascular mortality among individuals with diabetes (van Dooren et al., 2013). In addition, studies have shown that diabetes itself is a significant risk factor for stroke. The Northern Manhattan Study demonstrated that each additional year of diabetes duration increases the risk of ischemic stroke by 3%, with risk tripling in individuals living with diabetes for over ten years (Banerjee et al., 2012). Furthermore, multimorbidity has been consistently associated with worse post-stroke outcomes and increased mortality in large-scale cohort studies (Gallacher et al., 2018). Gallacher analysed data from 8,751 UK Biobank participants with stroke or transient ischaemic attack (TIA) and found that over 85% had at least one additional long-term condition. Increasing multimorbidity was associated with higher all-cause mortality, with diabetes, depression, cancer, and coronary heart disease among the conditions significantly linked to increased risk. These results underscore the importance of incorporating multimorbidity into stroke research and clinical guidelines.

In addition to clinical risk factors, biomarker research has identified elevated levels of glial fibrillary acidic protein (GFAP), neurofilament light chain (NfL), and phosphorylated tau (p-tau181 and p-tau217) as markers of neuroinflammation and white-matter injury (Sanchez et al., 2025). These markers may link neurovascular damage with neurodegenerative processes.

Despite these findings, the interactive effects of depression and diabetes on stroke risk

and outcomes remain undercharacterized, particularly in studies that integrate longitudinal data and advanced modeling strategies. To address these gaps, this study aims to first use the Atherosclerosis Risk in Communities (ARIC) longitudinal cohort and Cox proportional hazards models to quantify the joint effect of depression and diabetes on time to stroke occurrence, providing interpretable hazard ratios for relative risk. Second, we will develop and validate machine learning models, including XGBoost, random forests, and neural networks, to evaluate whether incorporating comorbidity information improves personalized stroke risk prediction. By integrating these approaches, our study seeks both to clarify the association between comorbidity and stroke risk and to enhance predictive stratification for high-risk individuals.

3 Methods

3.1 Data Source

The Atherosclerosis Risk in Communities (ARIC) Study is a prospective, community-based cohort established to investigate the causes and clinical consequences of atherosclerosis (The ARIC Investigators, 1989). Between 1987 and 1989, the study enrolled 15,792 adults aged 45–64 years from four U.S. communities: Forsyth County, North Carolina; Jackson, Mississippi; suburban Minneapolis, Minnesota; and Washington County, Maryland. Recruitment began in 1987, with baseline examinations completed by 1989.

Participants have been followed longitudinally through repeated in-person examinations and annual (semiannual since 2012) telephone interviews to ascertain cardiovascular events, comorbid conditions, hospitalizations, and mortality. As of 2020, eight examination visits have been completed: Visit 1 (1987–1989), Visit 2 (1990–1992), Visit 3 (1993–1995), Visit 4 (1996–1998), Visit 5 (2011–2013), Visit 6 (2016–2017), Visit 7 (2018–2019), and Visit 8 (2020).

The ARIC Study is supported by the National Heart, Lung, and Blood Institute (NHLBI) and approved by the institutional review boards of all participating institutions. Written informed consent was obtained from all participants at each study visit(The ARIC Investigators, 1989; Wright et al., 2021).

3.2 Study Population and Inclusion/Exclusion Criteria

Participants who attended Visit 5 (2011–2013) of the Atherosclerosis Risk in Communities (ARIC) Study (n = 6,538) were eligible for inclusion. This Study used Visit 5 as the baseline. Prevalent ischemic stroke or transient ischemic attack (TIA) at visit 5 was determined base on self-reported stroke history and adjudicated surveillance events. Participants identified as having prevalent ischemic stroke or transient ischemic attack at visit 5 were excluded (n = 209). Participants were excluded if they had missing data on diabetes status (n = 262). Participants missing depression assessment based on the self-reported depression scale were excluded (n = 139). Participants with missing prior stroke/TIA status (n = 11), or other key baseline covariates (n = 445) were also excluded. In addition, individuals who self identified with racial groups other than Black or White were excluded due to small sample sizes (n = 18). The final analytic sample comprised 5,459 participants (Figure 3.2).Based on Visit 5 assessments, participants were categorized into four exposure groups: (1) no diabetes and no depression, (2) diabetes only, (3) depression only, and (4) both diabetes and depression.

3.3 Exposure definition

3.3.1 Diabetes

Diabetes was defined as meeting at least one of the following criteria: self-reported physician diagnosis, self-reported use of diabetes medications, non-fasting blood glucose level $\geq$200 mg/dL, fasting blood glucose (FBG) $\geq$126 mg/dL, or hemoglobin A1c (HbA1c) $\geq$6.5%. Among individuals identified as having diabetes, glycemic control was further categorized as controlled (HbA1c $<$ 7%) or uncontrolled (HbA1c $\geq$7%), consistent with glycemic targets recommended by the American Diabetes Association (American Diabetes Association, 2021).

3.3.2 Depression

Depressive symptoms at Visit 5 were assessed using the 11-item version of the Center for Epidemiologic Studies Depression Scale (CES-D), a validated short form of the original 20-item scale developed for use in older adult populations (Kohout et al., 1993). The CES-D score was analyzed both as a continuous measure to capture the severity of depressive symptoms and as a binary indicator of clinically elevated symptoms. Based on prior validation research, a cutoff score of $\geq$9 was used to define clinically significant depressive symptoms (Takeshita et al., 2002).

3.4 Outcome definition: Stroke

The primary outcome was incident stroke, defined as the first definite or probable stroke event (of any type) occurring after Visit 5 among participants without prior stroke at baseline. Stroke events were ascertained using ARIC's active surveillance system, which included annual or semiannual participant interviews, review of hospital discharge summaries and ICD-9-CM codes (430–438), and keyword searches in medical records (e.g., "stroke," "aphasia," "cerebrovascular disease"). Stroke hospitalizations were eligible for validation if any cerebrovascular ICD codes were recorded or relevant keywords were identified in discharge summaries or clinical notes. Detailed medical record abstraction was performed, and diagnoses were adjudicated by trained physicians based on modified National Survey of Stroke criteria.

A stroke was classified as definite or probable if there was evidence of a sudden or rapid onset of neurological symptoms lasting >24 hours or leading to death, in the absence of a non-stroke cause (e.g., trauma, tumor, metabolic coma). Diagnostic confirmation was based on clinical presentation, imaging (CT/MRI), and physician review. The final classification incorporated both computer-generated diagnoses and physician consensus. Disagreements were resolved by a second reviewer (Rosamond et al., 1999).

3.5 Covariates

All covariates were assessed at Visit 5 (2011–2013) unless otherwise specified (Graff-Radford et al., 2017).

3.5.1 Sociodemographic Variables

Age Age was calculated from the participant's date of birth and the date of Visit 5. The mean age of participants at Visit 5 was approximately 75 $\pm$ 5 years.

Gender Gender was self-reported as male or female.

Race Race was self-reported and categorized as Black or White. Participants of other racial groups (e.g., Asian, Native American) were excluded due to small sample sizes.

Center Center refers to the ARIC field centers where participants were recruited and examined. These include: Washington County, Maryland; Forsyth County, North Carolina; Jackson, Mississippi; and suburbs of Minneapolis, Minnesota.

3.5.2 Lifestyle Variables

Smoking status Smoking status was self-reported at Visit 5 using ARIC study protocols and categorized as never, former, or current smoker. In the ARIC questionnaire, participants who reported smoking fewer than 400 cigarettes in their lifetime were classified as never smokers, consistent with ARIC definitions. Some prior literature had used a $\geq$100 cigarette threshold to define ever smoking(Howard et al., 1998), however, this study adheres to the original ARIC classification. Former smokers had smoked at least 400 cigarettes in their lifetime but had quit smoking at the time of the interview. Current smokers were currently smoking at the time of the interview, either every day or on some days.

Drinker status Drinker status was self reported using the ARIC Alcohol and Smoking Form. Participants were categorized as never, former, or current drinkers.

3.5.3 Clinical and Anthropometric Variables

Body Mass Index (BMI) BMI was calculated as weight in kilograms divided by height in meters squared (kg/m^2). Participants were categorized as underweight (< 18.5), normal weight (18.5–24.9), overweight (25–29.9), or obese (≥ 30).

Hypertension Hypertension was defined as systolic blood pressure (BP) $\geq$140 mmHg, diastolic BP $\geq$90 mmHg, or current use of antihypertensive medications. Systolic and diastolic BP were measured during the physical examination.

Antihypertensive Medication Use of antihypertensive medications was based on self-reported medication use during the past four weeks prior to Visit 5. Participants were instructed to bring all prescription and over-the-counter medications taken within the two weeks prior to the clinic visit, including vitamins and supplements, in their original containers ("ARIC Visit5 Medication Instruction Sheet", 2011).

3.5.4 Medical history variables

Previous CHD Previous Coronary Heart Disease (CHD) was defined as death from CHD and definite or probable fatal and nonfatal myocardial infarction (MI).

Previous Heart Failure Previous Heart Failure (HF) was defined as a physician report or hospitalization records with ICD-9 code 428.x in the primary discharge position (Rosamond et al., 1999).

3.6 Statistical Analysis

A two stage analysis was conducted to evaluate the association between comorbidity of depression and diabetes with incident stroke, using both traditional Cox proportional

hazards regression and machine learning prediction models. Additionally, logistic regression with stepwise selection was used to supplement the predictive findings of the machine learning model.

3.6.1 Cox Proportional Hazards Model

The Cox proportional hazards model was used to estimate the association between comorbidity status and risk of incident stroke. The Cox model is expressed by the hazard function denoted by h(t). It can be expressed as follow (Cox, 1972):

$$h(t|X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \tag{1}$$

where:

- $h(t|X)$: hazard function at time t for an individual with covariate vector X
- $h_0(t)$: baseline hazard function
- $\beta_1, \beta_2, \ldots, \beta_p$: regression coefficients
- $X_1, X_2, \ldots, X_p$: covariates included in the model

The exposure variable was a four level categorical variable indicating comorbidity status: No diabetes and no depression (reference group), Diabetes only, Depression only, Both diabetes and depression. Hazard ratios and 95% confidence intervals were estimated in three models: Unadjusted model, Adjusted for some variables, Fully adjusted for all variables. Kaplan–Meier survival curves were generated to compare stroke-free survival across comorbidity groups. Statistical differences in survival curves were evaluated using the log-rank test, Wilcoxon test, and likelihood ratio test. To assess the proportional hazards assumption of the Cox models, log(-log) survival curves were plotted against log(time). A two-sided p-value < 0.05 was considered statistically significant.

3.6.2 Machine Learning Prediction Model

To complement traditional Cox regression analysis and explore the predictive utility of depression-diabetes multimorbidity, a machine learning model was developed to classify individuals at high risk of stroke. Participants with complete outcome and covariate data were included. The outcome variable was incident stroke. Predictor variables included sociodemographic, lifestyle, clinical, and medical history factors from Visit 5. Categorical variables were converted into dummy variables. Additional derived features were engineered. The dataset was split into training (80%) and testing (20%) sets. Given the low prevalence of incident stroke, Borderline-SMOTE was applied to oversample the minority class in the training set. All features were scaled using StandardScaler. XGBoostClassifier, a gradient boosting decision tree algorithm, with hyperparameters optimized through grid search and 3-fold cross-validation was used. The XGBoost model minimizes the following regularized logistic loss function for binary classification (T. Chen & Guestrin, 2016):

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3}$$

- $\mathcal{L}(\phi)$: total objective function to be minimized
- $l(y_i, \hat{y}_i)$: logistic loss between the observed outcome y_i and predicted probability $\hat{y}_i$
- f_k: the k-th regression tree in the ensemble
- $\Omega(f_k)$: regularization term penalizing the complexity of tree f_k
- T: number of leaves in the decision tree
- w: vector of scores on the leaves
- γ: penalty for each additional leaf node
- λ: L2 regularization parameter on leaf weights

Model performance was evaluated on the testing set using Area Under the Receiver Operating Characteristic Curve, Area Under the Precision-Recall Curve, Classification report ,Confusion matrix, and Threshold optimization to achieve recall $\geq 60\%$, Model interpretability was further explored using feature importance plots.

3.6.3 Logistic Regression Model

We additionally conducted a logistic regression analysis with stepwise variable selection to explore the predictive ability of depression-diabetes multimorbidity and other covariates for incident stroke. The logistic regression model has the following form:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{4}$$

where:

- p: the predicted probability of incident stroke
- β_0: intercept term
- $\beta_1, \beta_2, \ldots, \beta_k$: regression coefficients
- $X_1, X_2, \ldots, X_k$: covariates selected via stepwise variable selection

Stepwise selection was based on the Akaike Information Criterion (AIC), which evaluates model fit while penalizing model complexity(Akaike, 1974).

$$\text{AIC} = 2k - 2\ln(L) \tag{5}$$

where:

- k: the number of model parameters
- L: the maximum likelihood of the model

The final model's predictive performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC). An AUC value closer to 1 indicates better discrimination.

By integrating traditional regression models (Cox and logistic) with machine learning approaches, this study aimed to both quantify the relative risk of stroke associated with depression-diabetes multimorbidity and evaluate its predictive value.

3.7 Variables included in each model

3.7.1 Cox Proportional Hazards Model

Exposure: Comorbidity status of depression and diabetes, modeled as a categorical variable with four levels:
0 = neither condition (reference), 1 = diabetes only, 2 = depression only, 3 = both conditions.

Outcome: Incident stroke, which is time to stroke (in days) paired with stroke status (where 1 = stroke event, and 0 = censored (no event))

Table 1: Covariates included in Each Cox Proportional Hazards Model

Model	**Covariates**
Model 1 (Unadjusted)	None
Model 2 (Adjusted)	Age Gender (reference: Female) Race (reference: White)
Model 3 (Fully Adjusted)	Age Gender (reference: Female) Race (reference: White) Smoking status (reference: Never) Drinking status (reference: Never) Body mass index Hypertension (reference: No) Antihypertensive medication use (reference: No) History of heart failure (reference: No) History of coronary heart disease (reference: No)

3.7.2 Machine Learning Prediction Model

The outcome variable was incident stroke (binary). Features included the following: Age, Gender, Race, Smoking status, Drinking status ,Body Mass Index, Hypertension status, Antihypertensive medication use, History of coronary heart disease, History of heart failure, Comorbidity status of depression and diabetes

3.7.3 Logistic Regression Model

To examine whether depression and diabetes exerts a synergistic effect on stroke risk, an interaction term was created, defined as the product of depression status and diabetes status. This variable equals 1 only when both conditions are present and 0 otherwise. This interaction was included with other covariates in a multivariable logistic regression model using stepwise variable selection. Other covariates include: Depression status, Diabetes status, Age, Gender, Race, Smoking status, Drinking status ,Body Mass Index, Hypertension status, Antihypertensive medication use, History of coronary heart disease and History of heart failure. The outcome is stroke.

4 Results

4.1 Baseline Characteristics

Among the 5459 participants in the study, the mean age was 75.4 years, and 58.3 percent were female. 3527 of 5459 participants had no diabetes or depression, 1594 had diabetes only, 185 had depression only, with 153 had both conditions. Among the four exposure groups, individuals with both depression and diabetes had the highest mean BMI, and the highest proportions of females, Black participants, hypertension, use of antihypertensive medication, and prevalent heart failure. Among all participants, 4.3% (n=233) experienced an incident stroke during follow-up. The stroke risk was highest in participants with both diabetes and depression (7.8%). These results suggest a possible synergistic effect of diabetes and depression on stroke risk.

Table 2: Baseline Characteristics of Participants by Comorbidity Status

Variable	**Total Cohort** (n=5459)	**No Diabetes/ Depression** (n=3527)	**Diabetes Only** (n=1594)	**Depression Only** (n=185)	**Diabetes and Depression** (n=153)
Continuous variables mean (SD)					
BMI(kg/m^2)	28.75 (5.69)	27.66 (5.23)	30.70 (5.76)	28.89 (5.88)	33.39 (7.06)
CES-D Score	3.04 (2.96)	2.42 (2.14)	2.76 (2.25)	10.86 (1.97)	10.84 (2.07)
Age(years)	75.42 (5.08)	75.39 (5.06)	75.44 (5.04)	76.19 (5.79)	74.97 (5.30)
Systolic BP(mmHg)	130.12 (18.04)	129.94 (17.72)	130.37 (18.70)	130.71 (17.28)	130.91 (19.25)
Diastolic BP(mmHg)	66.22 (10.73)	66.65 (10.70)	65.08 (10.67)	67.67 (10.32)	66.31 (11.50)

Continued on next page

Table 2 (continued)

Variable	Total Cohort	No Diabetes/ Depression	Diabetes Only	Depression Only	Diabetes and Depression
Categorical variables n (%)					
Center					
Forsyth	1151 (21.08%)	788 (22.34%)	293 (18.38%)	44 (23.78%)	26 (16.99%)
Jackson	1126 (20.63%)	587 (16.64%)	425 (26.66%)	56 (30.27%)	58 (37.91%)
Minneapolis	1699 (31.12%)	1250 (35.44%)	405 (25.41%)	33 (17.84%)	11 (7.19%)
Washington	1483 (27.17%)	902 (25.57%)	471 (29.55%)	52 (28.11%)	58 (37.91%)
Gender					
Female	3182 (58.29%)	2112 (59.88%)	843 (52.89%)	119 (64.32%)	108 (70.59%)
Male	2277 (41.71%)	1415 (40.12%)	751 (47.11%)	66 (35.68%)	45 (29.41%)
Race					
Black	1226 (22.46%)	638 (18.09%)	467 (29.30%)	60 (32.43%)	61 (39.87%)
White	4233 (77.54%)	2889 (81.91%)	1127 (70.70%)	125 (67.57%)	92 (60.13%)
Drinking Status					
Current	2721 (49.84%)	1936 (54.89%)	670 (42.03%)	76 (41.08%)	39 (25.49%)
Former	1570 (28.76%)	893 (25.32%)	552 (34.63%)	61 (32.97%)	64 (41.83%)
Never	1168 (21.40%)	698 (19.79%)	372 (23.34%)	48 (25.95%)	50 (32.68%)
Smoking Status					

Continued on next page

Table 2 (continued)

Variable	Total Cohort	No Diabetes/ Depression	Diabetes Only	Depression Only	Diabetes and Depression
Current	320 (5.86%)	209 (5.93%)	77 (4.83%)	21 (11.35%)	13 (8.50%)
Former	2650 (48.54%)	1685 (47.77%)	807 (50.63%)	93 (50.27%)	65 (42.48%)
Never	2173 (39.81%)	1440 (40.83%)	615 (38.58%)	56 (30.27%)	62 (40.52%)
Unknown	316 (5.79%)	193 (5.47%)	95 (5.96%)	15 (8.11%)	13 (8.50%)
Hypertension Status					
No	1412 (25.87%)	1105 (31.33%)	232 (14.55%)	57 (30.81%)	18 (11.76%)
Yes	4047 (74.13%)	2422 (68.67%)	1362 (85.45%)	128 (69.19%)	135 (88.24%)
Prevalent Heart Failure					
No	4792 (87.78%)	3214 (91.13%)	1313 (82.37%)	155 (83.78%)	110 (71.90%)
Yes	667 (12.22%)	313 (8.87%)	281 (17.63%)	30 (16.22%)	43 (28.10%)
Prevalent CHD					
No	4642 (85.03%)	3082 (87.38%)	1274 (79.92%)	160 (86.49%)	126 (82.35%)
Yes	817 (14.97%)	445 (12.62%)	320 (20.08%)	25 (13.51%)	27 (17.65%)
Incident Stroke					
No	5226 (95.73%)	3391 (96.40%)	1517 (95.17%)	177 (95.68%)	141 (92.16%)
Yes	233 (4.27%)	136 (3.86%)	77 (4.83%)	8 (4.32%)	12 (7.84%)

4.2 Cox Proportional Hazards Model

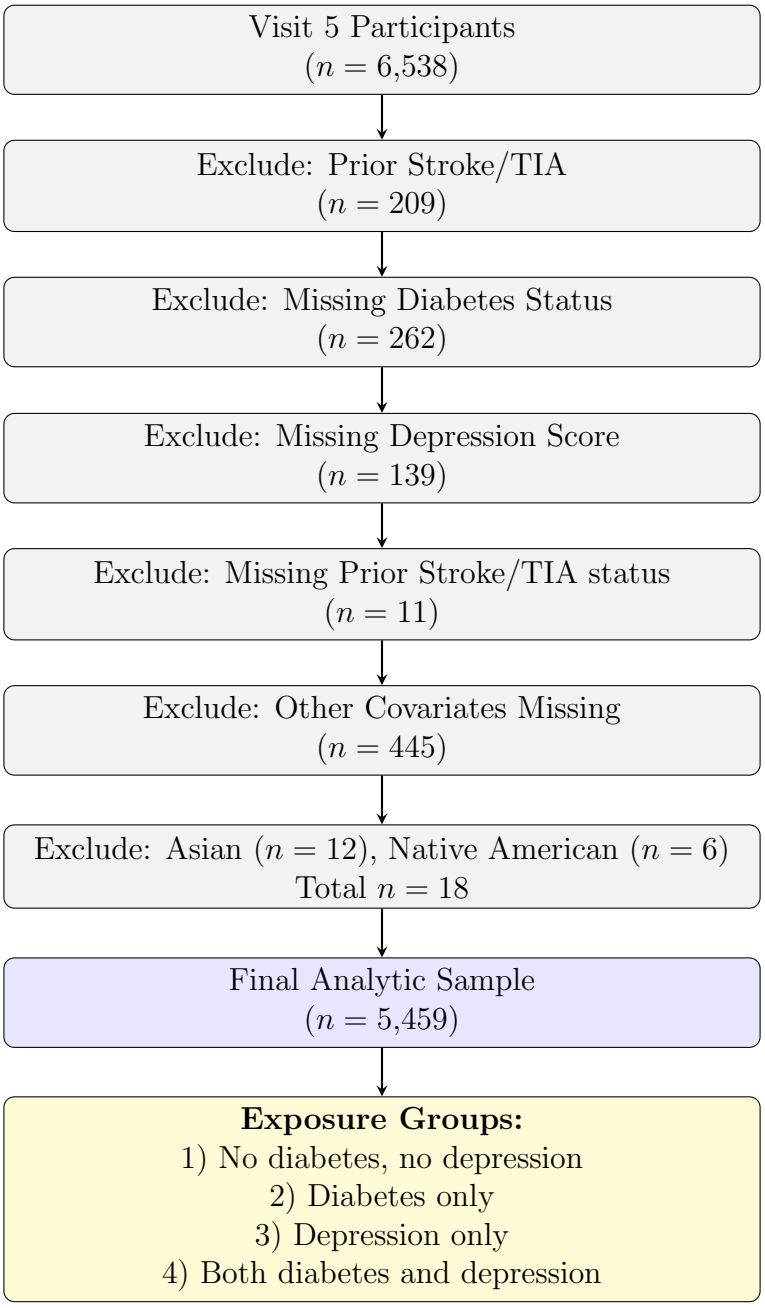

Figure 1: Study Population Selection Diagram

Table 3: Hazard Ratios (HR) and 95% Confidence Intervals for Incident Stroke by Exposure Group

Exposure	**Stroke events (n)**	**Model 1 HR (95% CI)**	**Model 2 HR (95% CI)**	**Model 3 HR (95% CI)**
No Diabetes or Depression	136	1.00	1.00	1.00
Diabetes only	77	1.367 (1.034–1.809)	1.335 (1.007–1.770)	1.215 (0.906–1.629)
Depression only	8	1.310 (0.642–2.673)	1.215 (0.595–2.483)	1.120 (0.546–2.296)
Diabetes and Depression	12	2.379 (1.318–4.295)	2.359 (1.302–4.274)	1.946 (1.051–3.601)

Model 1: Unadjusted According to the unadjusted Cox proportional hazards model, individuals with diabetes and depression multimorbidity had the highest risk of incident stroke. Compared to those without diabetes or depression, individuals with diabetes only had a 37% higher risk of stroke (HR = 1.367, 95% CI: 1.034–1.809), and those with depression had a 31% higher risk (HR = 1.310, 95% CI: 0.642–2.673). Individuals with both diabetes and depression had a 132% higher risk of stroke (HR = 2.379, 95% CI: 1.318–4.295) compared to reference group. The confidence interval does not include 1, indicating that this association is statistically significant. However, the small event counts in Depression only group and Diabetes and Depression group yield wider confidence intervals. These findings highlight the increased stroke risk associated with depression and diabetes, especially when both conditions are present.

Model 2: Adjusted for Age Gender Race Adjusting for age, sex and race in model 2, the association remained significant for the multimorbidity group (HR=2.359, 95% CI: 1.302–4.274) with a hazard ratio comparable to that in unadjusted model. Multimorbidity status still plays an important role in predicting stroke.

Model 3: Fully adjusted In the fully adjusted model, the hazard ratio for the multimorbidity group remained high (HR=1.946, 95% CI: 1.051–4.274). The association was statistically significant,indicating the effect of multimorbidity on stroke risk.

Kaplan-Meier Survival Curve Analysis and Log-Rank Tests:

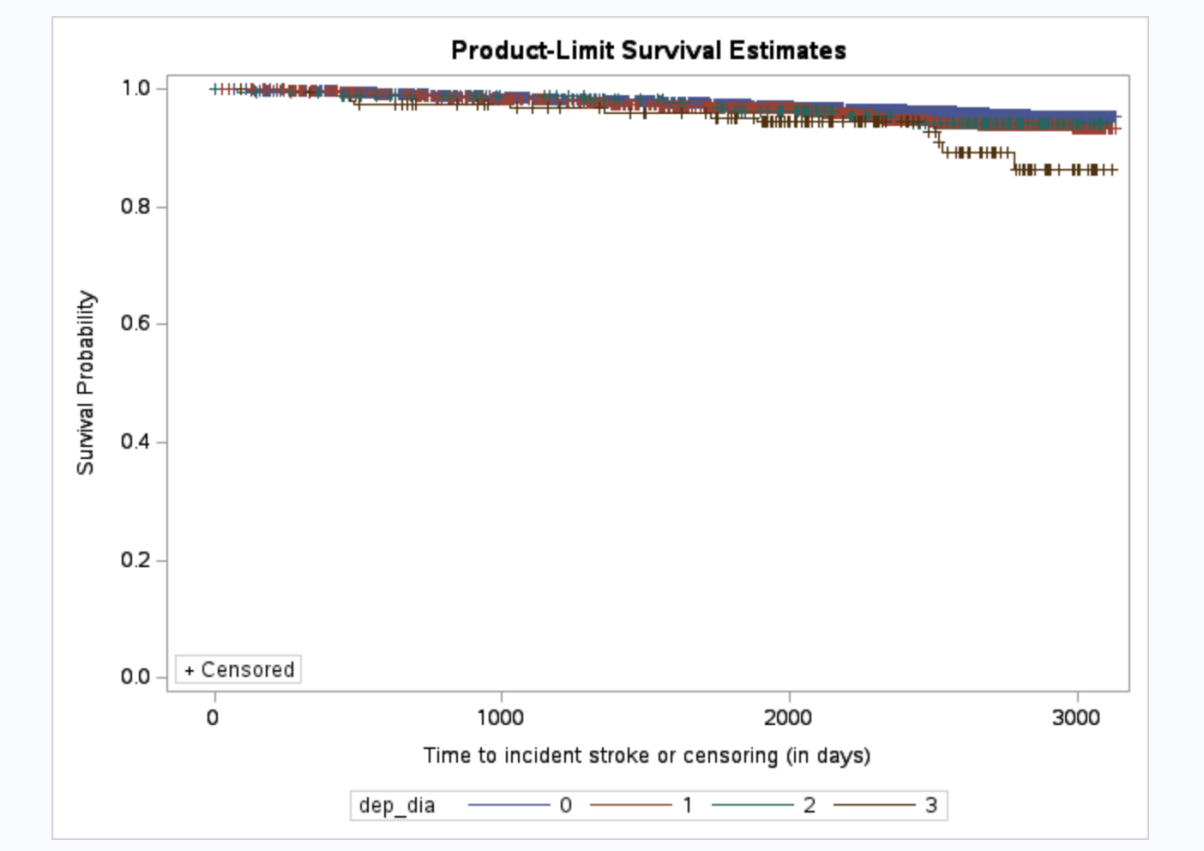

Figure 2: Kaplan–Meier survival curves by comorbidity group

Figure 2 displays the Kaplan–Meier survival curves stratified by comorbidity status. Individuals with both diabetes and depression consistently exhibited the lowest survival probabilities over time, reflecting the highest risk of incident stroke. In contrast, participants without diabetes or depression had the highest survival probabilities, suggesting the lowest stroke risk. These visual patterns are consistent with the findings from the Cox regression models, which revealed a stepwise increase in hazard ratios with increasing comorbidity burden.

Statistical tests confirmed the differences in survival across comorbidity groups, with significant results from the log-rank test ($\chi^2 = 11.74$, $p = 0.0083$), the Wilcoxon test ($\chi^2 = 8.75$, $p = 0.0329$), and the likelihood ratio test ($-2 \log L$, $\chi^2 = 9.33$, $p = 0.0252$). These results underscore the prognostic importance of comorbidity of depression and diabetes in predicting stroke risk.

Log(-log) Survival Plot: Assessment of Proportional Hazards Assumption

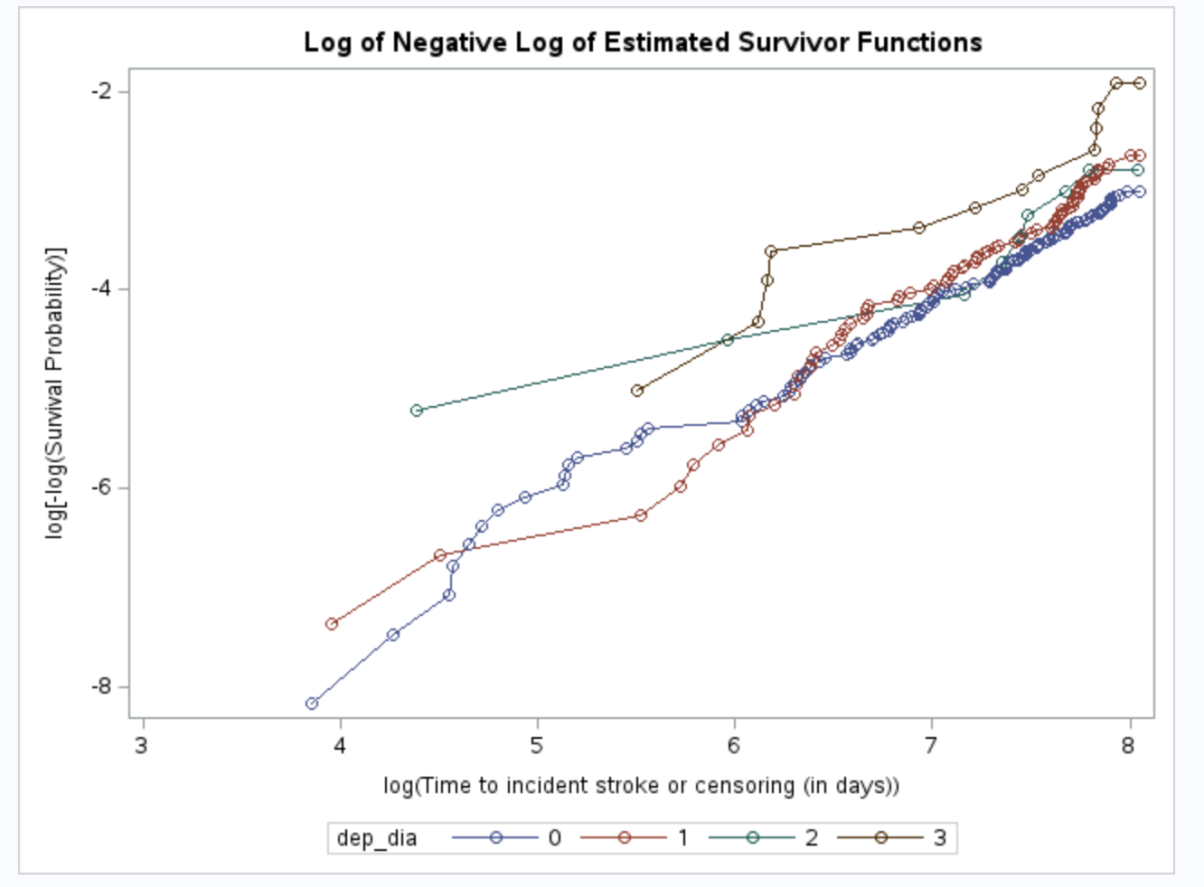

Figure 3: Log(-log) survival curves by comorbidity group

Figure 3 presents the log(-log) survival curves plotted against log(time) for each comorbidity group. Ideally, if the assumption holds, the curves for different groups should be approximately parallel. In this case, some deviations are observed, particularly for the multimorbidity group. These deviations may be attributed to smaller event counts and increased right-censoring in certain strata. While the curves were generally aligned, the proportional hazards assumption was considered acceptable for the present analysis.

4.3 Machine Learning model: Prediction of stroke risk

To investigate the predictive potential of depression-diabetes multimorbidity on stroke risk, a machine learning model was developed.

Table 4: Classification Metrics for Stroke Prediction at Two Thresholds

Threshold = 0.5			
Class	**Precision**	**Recall**	**F1 Score**
No Stroke	0.96	0.61	0.74
Stroke	0.04	0.40	0.08
Accuracy = 0.60			
Threshold = 0.258			
Class	**Precision**	**Recall**	**F1 Score**
No Stroke	0.96	0.45	0.61
Stroke	0.05	0.62	0.09
Accuracy = 0.45			

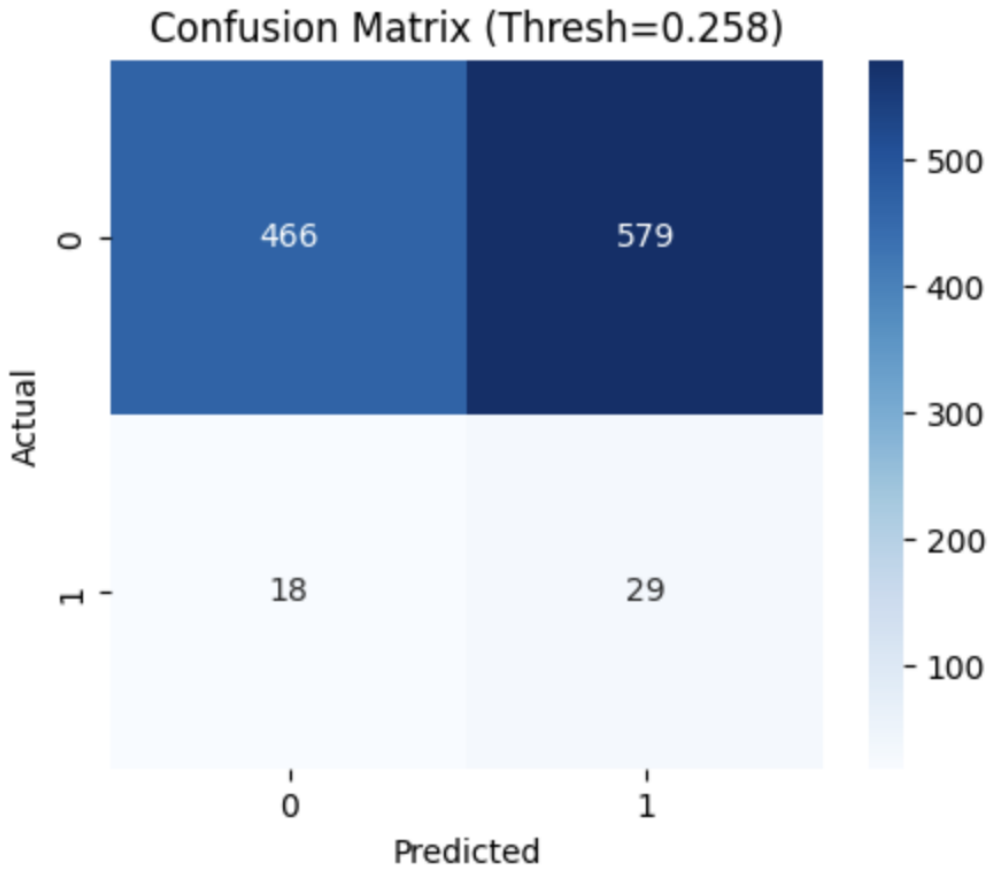

Figure 4: Confusion Matrix

This figure presents a Confusion Matrix, which summarizes the performance of a binary classification model in predicting stroke at a threshold of 0.258. The matrix shows:

- **True Negatives (TN)**: 466 individuals correctly identified as not having a stroke.
- **False Positives (FP)**: 579 individuals incorrectly predicted to have a stroke.
- **False Negatives (FN)**: 18 individuals incorrectly predicted as stroke-free.
- **True Positives (TP)**: 29 individuals correctly identified as having a stroke.

Table 4 shows that the model achieved high precision for stroke free individuals, indicating that most predicted non-stroke cases were accurate. However, recall for stroke cases was low, reflecting underdetection of the minority class when using the default classification threshold of 0.5. In After adjusting the threshold, recall for stroke cases were improved to 0.62, enhancing the model's potential utility for screening purposes. However, this improvement came at the cost of reduced precision, leading to a higher rate of false positives, which is a trade off.

While the model captured a signal from depression and diabetes multimorbidity with other covariates, its overall performance in identifying high and low risk individuals remained limited. These results indicate that depression and diabetes comorbidity contributes to stroke risk stratification. However, the severe class imbalance and restricted predictor strength may have constrained the model's predictive capability.

4.4 Logistic Regression Model

Table 5: Summary of Stepwise Selection for Logistic Regression Predicting Stroke Risk

Step	Effect Entered	DF	Score Chi-Square	Pr > ChiSq
1	age	1	24.8979	<0.0001
2	Prevalent CHD	1	13.7432	0.0002
3	Depression × Diabetes	1	5.0107	0.0252
4	Hypertension	1	3.3129	0.0687

Table 5 summarizes the variables selected through stepwise logistic regression to predict incident stroke. Age was the strongest predictor, followed by prevalent coronary heart disease, which remained significant. Importantly, the interaction term between depression and diabetes was also retained in the model ($p = 0.0252$), suggesting a synergistic effect of the two conditions on stroke risk. Hypertension status showed a trend toward significance, indicating its potential relevance. The final stepwise logistic model achieved a c-statistic of 0.625 (95% CI 0.589–0.661). While a c statistics of 0.5 suggests no discrimination, and a c statistic of 1 means perfect discrimination, this model falls in the range of moderate but limited discrimination.

5 Discussion

In this community-based prospective cohort of the ARIC Study, we found that comorbidity of diabetes and depression was significantly associated with an increased risk of incident stroke. Individuals with both conditions had the highest stroke incidence rate (7.8%) and the highest hazard ratio across 4 exposure groups. In the fully adjusted Cox PH model, this comorbidity group exhibited a 95% higher risk of stroke compared to participants with neither condition. The Kaplan–Meier survival curves demonstrated the lowest stroke-free survival probability among individuals with both diabetes and depression, which aligned with the Cox regression findings. In addition, results from logistic regression with stepwise selection further supported these observations. The interaction term between diabetes and depression was statistically significant, suggesting a non-additive relationship between the two conditions. Together, these findings indicate that the comorbidity of diabetes and depression has a synergistic effect on stroke risk beyond their individual contributions.

This is the first study to evaluate the joint effect of diabetes and depression on incident stroke using Cox PH model for risk estimation and machine learning model for risk prediction, supplemented by stepwise logistic regression. Both diabetes and depression have been found independently associated with stroke. (Pan et al., 2011; Banerjee et al., 2012) However, limited research have been conducted on joint effects of diabetes and depression on incident stroke. One prior study by Ouk (Ouk et al., 2020) found out prestroke diabetes and depression were associated with increased risks of post stroke

outcomes (e.g.dementia/institutionalization), without addressing the impact on incident stroke.

Several biological mechanisms may explain the synergistic effects of diabetes and depression on stroke risk. Both diabetes and depression are associated with chronic low grade systemic inflammtion (Pradhan et al., 2001; Dantzer et al., 2008), a key contributor to atherosclerosis and cerebrovascular disease. Although diabetes and depression operate through distinct biological pathways, they converge on several shared pathophysiological endpoints related to stroke. In diabetes, hyperglycemia promotes oxidative stress, endothelial dysfunction, and vascular remodeling (Brownlee, 2005). In contrast, depression exerts its effects through neuroendocrine dysregulation, including hyperactivation of the hypothalamic-pituitary-adrenal (HPA) axis, elevated cortisol levels, and increased sympathetic nervous system activity (Gold et al., 2015; Musselman et al., 1998). Despite their different origins, both conditions contribute to endothelial injury, thrombophilia, and decreased vascular tone, which may act synergistically to accelerate the progression of ischemic cerebrovascular events. These provide a biological basis for the statistically significant interaction observed in this study. The joint effect of diabetes and depression may amplify cerebrovascular vulnerability beyond the sum of their individual effects, indicating a synergistic mechanism of increased stroke risk in comorbid individuals.

The machine learning model used in this study provided complementary insight into stroke risk prediction. The model achieved a recall of 0.62, suggesting a moderate ability to identify high risk individuals. Although the precision was low, the model may still be useful for preliminary screening in populations with comorbid diabetes and depression. The performance of the model is largely limited by class imbalance and the scope of predictor variables. We need to optimize and validate in external datasets.

These findings emphasize the importance of early identification and management of individuals with both depression and diabetes. As the comorbid population shows an increased risk in stroke, they should be considered as a priority population for primary stroke prevention. Early identification of depressive symptoms in diabetic patients, and vice versa, may help circumvent subsequent stroke through prompt intervention. Integrating mental health care into chronic disease management, especially in primary care settings, could help improve management of high risk individuals.

There are several limitations of our study. First, the number of the incident stroke events was relatively small. This results in class imbalance, limited statitical power, and wider confidence intervals for some estimates. Second, depression was assessed using CESD scale, a self-reported screening tool rather than clinical diagnosis, which may introduce misclassification bias. Third, the study population consisted primarily of older adults, which may limit the generalizability of the findings to younger population. Lastly, this is an observational study, we cannot establish causal relationship between diabetes, depression and stroke.

Despite these limitations, the strengths of the study include the use of a large, well characterized prospective cohort and the integration of both traditional epidemiological methods and machine learning approaches to evaluate hazard ratios and enhance risk prediction. This enables a more comprehensive understanding of how diabetes and depression contribute to stroke risk. Future research should include clinical diagnosis of depression and evaluate predictive models using real-world data across diverse populations. Expanding the sample size and making the phenotypic data more detailed may strengthen the predictive model performance, enhancing the applicability in a wider public health setting.

References

World Health Organization. (2025). *Stroke, cerebrovascular accident*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Centers for Disease Control and Prevention, N. (2024). National vital statistics system, mortality 2018–2023. multiple cause of death files, 2018–2023. cdc wonder online database. https://wonder.cdc.gov/ucd-icd10-expanded.html

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., Bittencourt, M. S., et al. (2022). Heart disease and stroke statistics—2022 update: A report from the american heart association. *Circulation*, *145*(8), e153–e639. https://doi.org/10.1161/CIR.0000000000001052

Chen, R., Ovbiagele, B., & Feng, W. (2016). Diabetes and stroke: Epidemiology, pathophysiology, pharmaceuticals and outcomes. *The American Journal of the Medical Sciences*, *351*(4), 380–386. https://doi.org/10.1016/j.amjms.2016.01.011

Emerging Risk Factors Collaboration. (2010). Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease. *The Lancet*, *375*(9733), 2215–2222. https://doi.org/10.1016/S0140-6736(10)60484-9

Pan, A., Keum, N., Okereke, O. I., Sun, Q., Kivimäki, M., Rubin, R. R., et al. (2011). Depression and risk of stroke morbidity and mortality: A meta-analysis and systematic review. *JAMA*, *306*(11), 1241–1249. https://doi.org/10.1001/jama.2011.1282

Sacco, S., Ornello, R., Ripa, P., Pistoia, F., Degan, D., Tiseo, C., et al. (2015). Metabolic syndrome and stroke: A meta-analysis of prospective studies. *Journal of the American Heart Association*, *4*(2), e001180. https://doi.org/10.1161/JAHA.114.001180

Ouk, M., Wu, C. Y., Colby-Milley, J., Fang, J., Zhou, L., Shah, B. R., et al. (2020). Depression and diabetes mellitus multimorbidity is associated with loss of independence and dementia poststroke. *Stroke*, *51*(12), 3658–3668. https://doi.org/10.1161/STROKEAHA.120.031068

van Dooren, F. E., Nefs, G., Schram, M. T., Verhey, F. R., Denollet, J., & Pouwer, F. (2013). Depression and risk of mortality in people with diabetes mellitus: A systematic review and meta-analysis. *PLoS ONE*, *8*(3), e57058. https://doi.org/10.1371/journal.pone.0057058

Banerjee, C., Moon, Y. P., Paik, M. C., Rundek, T., Mora-McLaughlin, C., Vieira, J. R., et al. (2012). Duration of diabetes and risk of ischemic stroke: The northern manhattan study. *Stroke*, *43*(5), 1212–1217. https://doi.org/10.1161/STROKEAHA.111.641381

Gallacher, K. I., McQueenie, R., Nicholl, B., Jani, B. D., Lee, D., & Mair, F. S. (2018). Risk factors and mortality associated with multimorbidity in people with stroke or transient ischaemic attack: A study of 8,751 uk biobank participants. *Journal of Comorbidity*, *8*, 1–10. https://doi.org/10.15256/joc.2018.8.140

Sanchez, E., Coughlan, G. T., Wilkinson, T., Ramirez, J., Mirza, S. S., Baril, A.-A., et al. (2025). Association of plasma biomarkers with longitudinal atrophy and microvascular burden on mri across neurodegenerative and cerebrovascular diseases [Advance online publication]. *Neurology*, *104*(7). https://www.neurology.org/

The ARIC Investigators. (1989). The atherosclerosis risk in communities (aric) study: Design and objectives. *American Journal of Epidemiology*, *129*(4), 687–702. https://pubmed.ncbi.nlm.nih.gov/2646917

Wright, J. D., Folsom, A. R., Coresh, J., Sharrett, A. R., Couper, D., Wagenknecht, L. E., & Heiss, G. (2021). The aric (atherosclerosis risk in communities) study:

Jacc focus seminar 3/8. *Journal of the American College of Cardiology*, *77*(23), 2939–2959. https://doi.org/10.1016/j.jacc.2021.04.035

American Diabetes Association. (2021). Glycemic targets: Standards of medical care in diabetes—2022. *Diabetes Care*, *44*(Suppl. 1), S73–S84. https://doi.org/10.2337/dc22-S006

Kohout, F. J., Berkman, L. F., Evans, D. A., & Cornoni-Huntley, J. (1993). Two shorter forms of the ces-d (center for epidemiological studies depression) depression symptoms index. *Journal of Aging and Health*, *5*(2), 179–193. https://doi.org/10.1177/089826439300500202

Takeshita, J., Masaki, K. H., Ahmed, I., Chiemi, K., Petrovitch, H., Ross, W., & White, L. R. (2002). Are depressive symptoms a risk factor for mortality in elderly japanese american men? the honolulu-asia aging study. *American Journal of Psychiatry*, *159*(7), 1127–1132. https://doi.org/10.1176/appi.ajp.159.7.1127

Rosamond, W. D., Folsom, A. R., Chambless, L. E., Wang, C.-P., McGovern, P. G., Howard, G., Cooper, L. S., Sorlie, P., & Prineas, R. J. (1999). Stroke incidence and survival among middle-aged adults: 9-year follow-up of the atherosclerosis risk in communities (aric) cohort. *Stroke*, *30*(4), 736–743. https://doi.org/10.1161/01.STR.30.4.736

Graff-Radford, J., Simino, J., Kantarci, K., Mosley, T. H. J., Griswold, M. E., Windham, B. G., & Knopman, D. S. (2017). Neuroimaging correlates of cerebral microbleeds: The aric study (atherosclerosis risk in communities). *Stroke*, *48*(11), 2964–2972. https://doi.org/10.1161/STROKEAHA.117.018336

Howard, G., Wagenknecht, L. E., Burke, G. L., et al. (1998). Cigarette smoking and progression of atherosclerosis: The atherosclerosis risk in communities (aric) study. *JAMA*, *279*(2), 119–124. https://doi.org/10.1001/jama.279.2.119

Aric visit5 medication instruction sheet. (2011). https://aric.cscc.unc.edu/aric9/sites/default/files/public/visitdocuments/v5/Manual%202%20Home%20and%20Field%20Center%20Procedures.pdf

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Pradhan, A. D., Manson, J. E., Rifai, N., Buring, J. E., & Ridker, P. M. (2001). C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA*, *286*(3), 327–334. https://doi.org/10.1001/jama.286.3.327

Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W., & Kelley, K. W. (2008). From inflammation to sickness and depression: When the immune system subjugates the brain. *Nature Reviews Neuroscience*, *9*(1), 46–56. https://doi.org/10.1038/nrn2297

Brownlee, M. (2005). The pathobiology of diabetic complications: A unifying mechanism. *Diabetes*, *54*(6), 1615–1625. https://doi.org/10.2337/diabetes.54.6.1615

Gold, P. W., Machado-Vieira, R., & Pavlatou, M. G. (2015). Clinical and biochemical manifestations of depression: Relation to the neurobiology of stress. *Neural Plast.*, *2015*, 581976. https://doi.org/10.1155/2015/581976

Musselman, D. L., Evans, D. L., & Nemeroff, C. B. (1998). The relationship of depression to cardiovascular disease: Epidemiology, biology, and treatment. *Archives of General Psychiatry*, *55*(7), 580–592. https://doi.org/10.1001/archpsyc.55.7.580