10-Year Probability of Stroke from the Framingham Heart Study

Analyst: Brenton Graham
Report generated: 11/23/2022

Introduction

The Framingham Heart Study is a multigenerational prospective cohort study that was initiated in 1948 to characterize risk factors associated with cardiovascular disease (CVD). Participants enrolled in the study have been examined biennially for clinical risk factors of CVD, including blood pressure, lung function, smoking history, and medication use, among other risk factors. CVD-related health outcomes such as incidence of stroke, myocardial infarction (i.e., heart attack), and heart failure have been tracked for all participants in the study along with time-to-event outcome measures. In this report, we focus on a subset of data from 4,434 participants of the Framingham Heart Study to identify sex-specific baseline risk factors associated with incidence of stroke. Further, we aim to estimate the 10-year probability of stroke associated with different risk profiles in this cohort. While we focus on baseline CVD risk factors for analysis in this report, we will also provide a high-level overview of how risk factors change over time (in relation to stroke incidence) and discuss how a time-varying covariate survival analysis might be of further interest.

Methods

Exploratory Data Analysis

Summary statistics and data visualizations were used to explore the relationship between baseline risk factors and 10-year incidence of stroke. Associations between continuous risk factors and 10-year incidence of stroke were visualized using density plots (Figure 1). Associations between categorical risk factors and 10-year incidence of stroke were visualized using compositional stacked bar plots (Figure 2). Subjects are stratified by sex and incidence of stroke during 10 years of follow-up in each figure.

Figure 1: Distributions of continuous CVD risk factors at baseline stratified by sex and 10-year incidence of stroke.

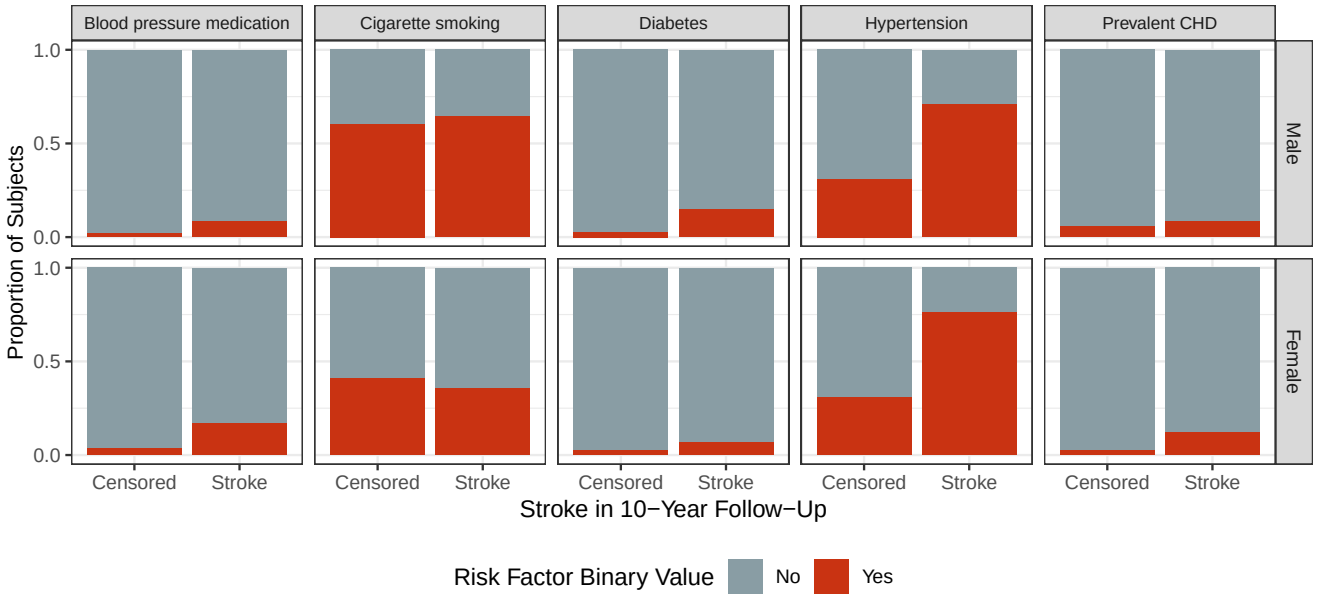

Figure 2: Comparison of categorical CVD risk factors at baseline stratified by sex and 10-year incidence of stroke.

Data Preparation

The original data set was filtered to contain only baseline CVD risk factors. Subjects with a history of stroke (32 subjects) were excluded from analysis. Time-to-event data were manipulated to reflect 10 years follow-up (rather than the 24 years of follow-up that are available in the data set) since we are interested in estimating 10-year probability of stroke. First, stroke events occurring after 10 years of follow-up were censored (i.e., set to non-events). Second, time-to-stroke measures were restricted to a maximum of 10 years. Death is treated as a censoring event throughout the analysis, whereby time-to-death is used as time-to-event follow-up time.

Variable Screening

Predictor variables were screened for redundancy, collinearity, missingness, and biological relevance. Predictor exclusion and exclusion rationale is described below.

- **Cigarette smoking.** The original data set included variables for both current smoking status and daily cigarette consumption. We prefer current smoking status as an indicator of cigarette smoking since there are missing data associated with daily cigarette consumption measurements.
- **Blood pressure.** We observe a strong correlation between systolic blood pressure (SBP) and diastolic blood pressure (DBP) in this data set (Pearson's $r = 0.79$), which could lead to issues with collinearity. This positive association is well reported throughout CVD literature[1]. Reports suggest SBP to be the better predictor of CVD-related events[2]. As such, DBP is excluded as a predictor in our models.
- **Cholesterol.** Total, high-density lipoprotein (HDL), and low-density lipoprotein (LDL) cholesterol variables were included in the original data set. However, HDL and LDL cholesterol measurements were not taken at baseline (i.e., these variables are completely missing for time point 1). Total cholesterol measurements become biologically obsolete when HDL and LDL cholesterol proportions are unknown. All cholesterol variables are therefore excluded from analysis.
- **Glucose.** A casual serum glucose variable is included in the data set. This variable is associated with more missingness than any other predictor in the data set (with 9% of participants missing measurements). The data set contains a separate variable which denotes diabetic and non-diabetic subjects and is partially derived

from the glucose measurements (i.e., subjects with a casual serum glucose level over 200 mg/dL are considered diabetic). The glucose variable is excluded from analysis since similar information is contained within the diabetes marker and since the diabetes marker is available for all participants. Glucose levels are confirmed to be significantly different between diabetic and non-diabetic participants ($p < 0.001$).

- **Prevalent CHD, MI and AP.** The data set contains separate markers for prevalent coronary heart disease (CHD), myocardial infarction (MI), and angina pectoris (AP). The prevalent CHD marker was found to encompass information from each of the other markers, and is used as a single indicator of pre-existing MI, AP, or coronary instability in our models.

Following variable screening we are left with 10 potential predictors of stroke incidence. These include age, SBP, cigarette smoking status, body mass index (BMI), heart rate, diabetes diagnosis, use of blood pressure medication, prevalent CHD, prevalent hypertension, and education status (binarized into less than high school degree or high school degree or more).

Missing Data

Missing data were present in 190 observations (4.3%) after variable selection; 83 of which were male (4.3% of all males) and 107 of which were female (4.2% of all females). Missing entries were determined to not be missing completely at random (MCAR) by Little's MCAR hypothesis test ($p < 0.001$). Nonetheless, we decided to move forward with a complete-case analysis due to time constraints. Limitations to this approach will be discussed in the limitations section.

Statistical Modeling

Statistical modeling was performed separately for male and female subjects in this study. Regularized Cox proportional hazards (CPH) models were fit using the LASSO (L_1) penalty for variable selection. Briefly, LASSO regression requires specification of a λ penalization hyperparameter which determines the extent of regularization. The optimal λ value was determined for each sex-specific model through K-fold cross-validation using 5-folds. The selected value of lambda corresponded to the minimal value of lambda (i.e., least penalization) that maximized Harrell's C-index. λ values of 0.00193 and 0.00241 were used for the male- and female-specific models, respectively. Separate CPH models were also fit without regularization as an exploratory way to assess shrinkage. Proportional hazards assumptions were confirmed for all Cox regression models using the `cox.zph` hypothesis test from the `survival` package in R.

Following model fitting, LASSO-based models were used to group individuals into three groups of stroke-related risk profiles, including low-, medium- and high-risk groups. Risk groups were characterized following a method proposed by Witten and Tibshirani (2010)[3]. Briefly, risk scores were computed for each subject from subject-specific covariates and the coefficient estimates from the fitted proportional hazards models. Subjects were then broken into equally-sized low-, medium- and high-risk groups based on risk score. Finally, Kaplan-Meier curves were fit to the risk group categories and p-values were used to measure how well risk groups are stratified in the model. A stringent significance value of $\alpha = 0.01$ is used throughout analysis to guard against Type-I errors that might arise from multiple comparisons.

Results

Table 1 describes the demographics and baseline risk factors of the study population that remained following data preparation, variable screening and complete-case selection. The table is stratified by sex and incidence of stroke to highlight the imbalanced proportion of subjects who experienced a stroke during the 10-year follow-up period. Certain risk factors are clearly imbalanced between the male and female populations, including cigarette smoking, which further supports the rationale of using a sex-stratified analysis.

Table 1: Cohort demographics stratified by sex and 10-year incidence of stroke

Risk Factor	Male		Female	
	Censored	Stroke	Censored	Stroke
n	1801	46	2307	58
Age (years)	49 [42, 57]	57 [50, 61]	49 [42, 56]	59 [53, 62]
Body Mass Index (kg/m^2)	26 [24, 28]	26 [24, 29]	25 [23, 28]	27 [23, 31]
Systolic Blood Pressure (mean mm Hg)	128 [118, 141]	144 [131, 173]	128 [116, 145]	157 [140, 177]
Heart Rate (beats/min)	75 [66, 80]	75 [67, 85]	75 [70, 85]	72 [68, 81]
Cigarette Smoker	1092 (60.6)	29 (63.0)	942 (40.8)	21 (36.2)
Diabetic	48 (2.7)	7 (15.2)	52 (2.3)	4 (6.9)
Using Blood Pressure Medication	36 (2.0)	3 (6.5)	83 (3.6)	10 (17.2)
Prevalent Coronary Heart Disease	105 (5.8)	4 (8.7)	55 (2.4)	7 (12.1)
Prevalent Hypertension	559 (31.0)	32 (69.6)	706 (30.6)	44 (75.9)
Education (Less than High School)	792 (44.0)	31 (67.4)	915 (39.7)	28 (48.3)

Note: Continuous variables are reported as median [IQR]. Categorical variables are reported as n, (%).

Hazard ratio estimates from both L_1-regularized and non-regularized Cox proportional hazards models are shown in Table 2. As mentioned, estimates and significance values from the non-regularized models were solely computed to provide a perspective on the shrinkage effect of the LASSO; these estimates will not be used for inference in this report. Standard tools of frequentist statistical inference including confidence intervals and p-values are not applicable (at least yet) to LASSO-based model estimates.

In the male-specific model, age, use of blood pressure medication, cigarette smoking, diabetes, education, prevalent hypertension, and systolic blood pressure are all identified as potentially useful predictors of 10-year stroke incidence. Diabetics are associated with a hazard ratio estimate of 4.18, indicating that males who are diagnosed as diabetic at baseline are 4.18 times more likely to experience a stroke over 10 years of follow-up than males who are not diagnosed as diabetic (controlling for all other covariates). In the female-specific model, age, use of blood pressure medication, cigarette smoking, diabetes, prevalent hypertension, and systolic blood pressure were again selected as potentially useful predictors of 10-year stroke incidence. Heart rate and prevalent CHD, which were not selected in the male-specific model, were additionally selected in this model. BMI was selected in neither circumstance. Age and systolic blood pressure are positively associated with 10-year hazard of stroke and return identical estimates in both the male and female models. C-statistics of 0.776 and 0.799 were achieved by the male- and female-specific models, respectively, indicating good predictive performance in each.

Table 2: Model estimates from Cox proportional hazards models with and without the LASSO

	Male Models			Female Models		
	LASSO-CPH	Cox PH		LASSO-CPH	Cox PH	
Risk Factor	HR	HR	p-value	HR	HR	p-value
Age	1.05	1.06 (1.02, 1.11)	0.005**	1.05	1.06 (1.02, 1.10)	0.004**
BMI	-	0.97 (0.89, 1.06)	0.563	-	1.00 (0.95, 1.05)	0.967
Using B.P. Meds	1.15	1.43 (0.42, 4.91)	0.570	1.46	1.48 (0.72, 3.05)	0.292
Cigarette Smoker	1.40	1.70 (0.88, 3.27)	0.112	1.11	1.51 (0.86, 2.65)	0.147
Diabetic	4.18	4.96 (2.12, 11.60)	<0.001***	1.43	1.88 (0.67, 5.27)	0.228
Education (< H.S.)	1.42	1.59 (0.83, 3.04)	0.161	-	0.95 (0.55, 1.64)	0.847
Heart Rate	-	1.00 (0.98, 1.02)	0.927	0.99	0.98 (0.96, 1.00)	0.123
Prev CHD	-	1.02 (0.35, 2.98)	0.965	1.70	1.83 (0.81, 4.16)	0.146
Prev Hypertension	1.87	2.18 (0.96, 4.97)	0.064	1.66	2.00 (0.92, 4.34)	0.081
Systolic B.P.	1.02	1.02 (1.01, 1.04)	0.008**	1.02	1.02 (1.01, 1.03)	<0.001***

Table 3: 10-year probability of stroke incidence stratified by sex and risk group

	Male 10-Yr Prob of Stroke (95% CI)	Female 10-Yr Prob of Stroke (95% CI)
Low-Risk	0.16% (0.00, 0.48)	0.38% (0.00, 0.82)
Medium-Risk	2.05% (0.89, 3.19)	1.43% (0.59, 2.27)
High-Risk	6.17% (4.10, 8.19)	5.95% (4.23, 7.64)

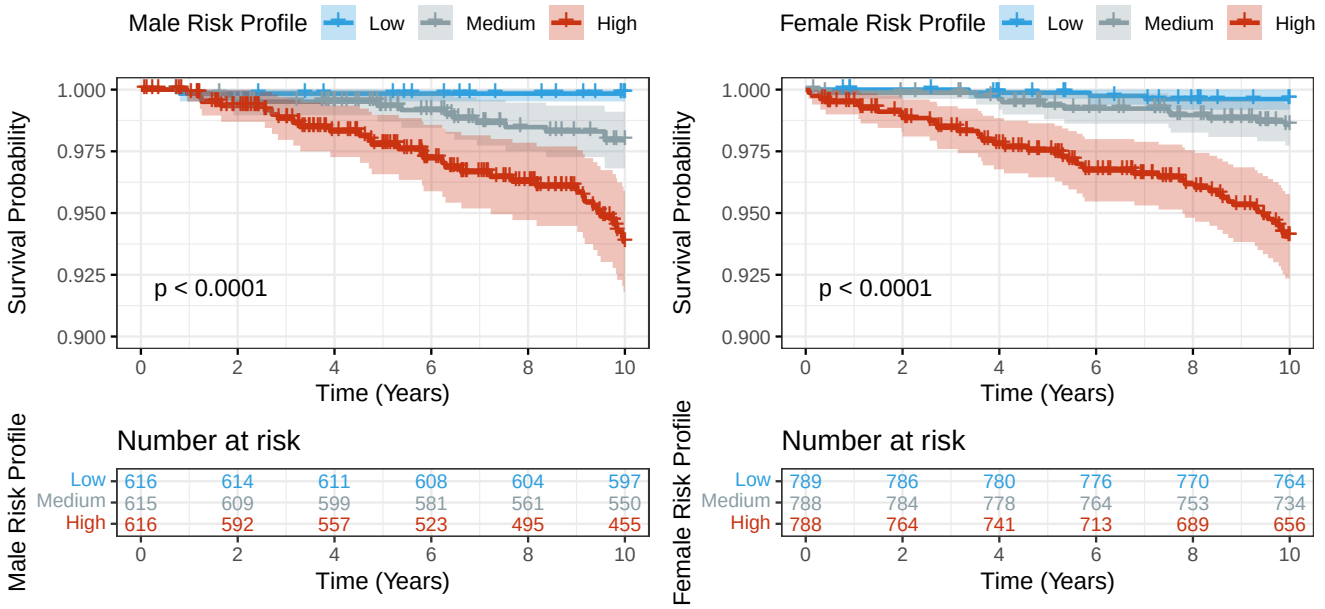

Figure 3: Kaplan-Meier 10-year survival curves stratified by risk profile.

As discussed, LASSO-based models were used categorize subjects into three risk group levels (low-, medium-, and high-risk of stroke incidence during 10-year follow-up) based on model-computed risk scores. Figure 3 shows the Kaplan-Meier 10-year survival curves that were fit and stratified using these risk set groups. Separation is observed between survival curves for each risk group ($p < 0.001$), with the lowest survival probability (i.e., highest probability of stroke) corresponding to the high-risk group. The 10-year probabilities of remaining uncensored and without stroke incidence for the low-, medium-, and high-risk groups are displayed in Table 3.

Discussion & Limitations

Results from this analysis suggest that 10-year probability of stroke incidence is associated with at least a few important CVD risk factors, and that these risk factors may vary between male and female subjects. The most influential risk factors of 10-year stroke appear to be baseline age, systolic blood pressure, and prevalent hypertension for both male and female subjects. Baseline diabetes status has a clear association with increased hazard of stroke incidence in male subjects.

While results are promising and appear to be consistent with the literature, limitations to this analysis are prevalent. First, a complete-case analysis was used for statistical modeling even though missing data were determined to not be MCAR. Future renditions of this study ought to use a multiple imputation approach for missing data handling and a sensitivity analysis should be performed to test for bias in our estimates. Second, the LASSO is known to produce biased estimators. It is therefore not recommended to depend on reported estimators for statistical inference in this report. Third, the LASSO is also known to produce unstable variable selection iteration-to-iteration and sample-to-sample. In other words, variables selected in our models are not guaranteed to be selected in models that are fit on a sub-sample of our population. A variable selection stability analysis, perhaps using a bootstrap approach, should

be explored to assess which variables are consistently selected as predictors of 10-year stroke incidence. Fourth, risk scores and corresponding risk profile groupings were computed using a LASSO-based Cox PH model that was trained on all available data. Therefore, the generalizability of these results should be questioned. Future analyses should use a training and validation set approach to test the generalizability of risk score predictions. Fifth, and finally, results from our study are restricted to baseline risk factors, while time-varying risk factors are likely more accurate predictors of stroke incidence. Further reports ought to explore the use of mixed effects Cox regression to assess this.

Citations

1. Sesso, H. D. *et al.* Systolic and diastolic blood pressure, pulse pressure, and mean arterial pressure as predictors of cardiovascular disease risk in men. *Hypertension* **36**, 801–807 (2000).
2. O'Rourke, M. Arterial stiffness, systolic blood pressure, and logical treatment of arterial hypertension. *Hypertension* **15**, 339–347 (1990).
3. Witten, D. M. & Tibshirani, R. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research* **19**, 29–51 (2010).