Project 3, Report

Joseph Froelicher

November 24, 2021

Contents

Introduction

Beginning in 1948, subjects were enrolled in a Heart disease study in the city of Framingham, Massachusetts. Our data come from a subset of that data consisting of the years 1956 - 1968. Participants were followed by hospital visits, participant contact and through death certificates for the occurrence of Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease. Using the 10-year probability of stroke based on different risk profiles, the goal is to identify any risk factors for stroke based on our data.

Using the provided Framingham heart data, we are interested in determining which risk factors are most associated with increased hazard of having a stroke, determining the 10 year risk for different risk profiles, and analyzing how much temporal change in the associated risk factors, and whether a more sophisticated temporal analysis is needed. Those associated risk factors include four binary categorical variables, presence of cardiovascular disease, smoking status, presence of diabetes, and treatment for hypertension, as well as four continuous risk factors, age, systolic blood pressure, body mass index, and cholesterol level.

Materials and Methods

Data

Our data consist of $4,434$ participants who's data were collected in 6 year periods from approximately 1956 - 1968. Each participant has anywhere from one to three observations, depending on how many examinations they attended ($11,627$ observations). During the cleaning process, 32 participants were dropped due to having a previous stroke in the first period, and another 127 participants were dropped for missing data, so the usable data has $4,275$ participants. The outcome for this analysis is Cerebral Hemorrhage (Stroke), defined by 10 year risk of stroke. The data include information about binary sex (M/F), age, blood pressure, smoking status, medication use, cholesterol, body mass index, and diabetes status. An additional variable was created to indicate whether subjects have previously acknowledged having any of: Angina Pectoris, coronary heart disease, and/or Myocardial infarction. We are interested in the time to event of Cerebral Hemorrhage (stroke), in the first 10 years. To measure this, two additional variables were created, one that measures the time to stroke or 10 years (whichever comes first), and one that indicates either a censoring (0), or a stroke event (1).

Descriptive Statistics

Reported in table 1 are the descriptive statistics for all of the non-missing data at hand. Categorical variables are reported as a percentage of the population, continuous variables are reported with means +/- standard deviation, and the time to stroke variable is reported with the median time +/- inter-quartile range. The table is stratified by stroke for each sex group (male and female).

Statistical Methods

A combination of statistical and exploratory approaches were used to address each of the aims mentioned previously. To assess the 10 year risk for different risk profiles, we are doing a survival analysis using Kaplan-Meier[1] curves to examine the risk for each of risk of interest. Each risk factor will have its own Kaplan-Meier curve, with binary risk factors shown with yes and no groups, and continuous variables have a curve for each quartile group (<25%, <50%, <75%, <100%). As well as building a table of survival probabilities for various risk profiles.

We are also fitting a Cox proportional-hazards[2] model for males and females separately to determine the risk factors that are associated with stroke. THe investigator was also interested in the interaction of systolic blood pressure and the use of blood pressure medication. Those risk factors will be selected using stepwise backwards selection (removal criterion: $p > 0.15$). To check the assumptions of Cox proportional-hazards models, the Schoenfeld residuals[3] will be checked.

And lastly a table of the risk factors over the first two periods will be provided to demonstrate any changes in the risk factors over the 10 years. Each of these analyses will be stratified by sex, to account for sex differences.

Results

As previously mention, the descriptive statistics are reported in table 1. In our data there were 2.46% who had a stroke in the 10 year period (n = 105).

Provided in figures 1 & 2 are the Kaplan-Meier curves for men and women respectively, for each of the risk factors of concern (listed above). On the right are the continuous risk factors, where participants were grouped based on quartiles, and on the left are binary risk factors. For females, presence of cardiovascular disease, use of hypertensive medication, presence of diabetes, old age, high blood pressure, and high body

mass index appear to be reducing the survival probability the most of all the risk factors. For males, use of hypertensive medication, presence of diabetes, increasing age, and high blood pressure appear to be reducing the survival probability the most of all the risk factors.

The results of the stepwise backwards model selection for the Cox proportional hazards models for both males and females are reported in table 2. For *males*, all of age, systolic blood pressure, smoking, and diabetes were found to be significant predictors. Age is accounting for a 6.6% (95% CI: [2.7%, 10.7%], p=0.001) increase in the hazard ratio for stroke. Systolic blood pressure is accounting for a 3.4% (95% CI: [2.2%, 4.6%], p<0.001) increase in the hazard ratio for stroke. Smoking is accounting for a 91.9% (95% CI: [3.4%, 256.4%], p=0.039) increase in the hazard ratio for stroke. Diabetes is accounting for a 379.9% (95% CI: 109.6%, 1098.9%], p<.001) increase in the hazard ratio for stroke. For *females*, only age and systolic blood pressure were selected as significant predictors the Cox proportional hazards model for stroke. Age is accounting for a 7.4% (95% CI: [3.4%, 11.6%], p<0.001) increase in the hazard ratio for stroke. Systolic blood pressure is accounting for a 2.6% (95% CI: [1.8%, 3.5%], p<0.001) increase in the hazard ratio for stroke.

Reported in table 3 are the 10-year risk profiles for various risk profiles of interest, based on a Cox proportional hazards model, for ages 55, 60, 65, 70, 75, 80, and 85 years. The risk factors (covariates) included in table 3 are all significant predictors, in addition to the investigator specified risk factors of interest, for our risk profiles.

Lastly, shown in table 4 are the changes from period 1 to period 2 of each of the risk factors of interest in this analysis. As to be expected, age is increasing, this is not of note. Continuous variables that are increasing over time that may be of note are: systolic blood pressure and cholesterol, and this is true for both males and females. The only categorical variable that changes over the two periods is smoking. There are more smokers in the data during period 2 than there are in period 1.

Discussion

Throughout this analysis, it was demonstrated that while graphically appearing significant, some risk factors were not necessarily statistically significant in the prediction of the hazard of stroke. The most notable risk factor for stroke for both males and females is age, and this was to be expected. With higher age, is higher risk for stroke. Some other variables, like systolic blood pressure and smoking were also statistically significant in the prediction of the hazard of stroke.

For future analyses, we may be interested in examining age or blood pressure as a confounder, or mediator, for these two variables may be interacting. Seeing as smoking was increasing over time and was also a significant predictor for males, further analyses may want to examine other coping mechanisms that persons who have experienced a stroke use. This is an interesting finding and warrants further analysis.

Lastly, in response to the investigators interest in the risk factors change over time, there was enough change over time to warrant examining a more sophisticated temporal model for these data. Some of the variables in period 2 are different than in period 1, and thus there may be need for better modeling.

Limitations

This study was limited primarily by missing data and partially as a result of that small sample sizes. The were low counts of strokes, and because the data were stratified by sex, the number of strokes was even smaller. Additionally, as was mentioned, some of the variables may be changing over time, and this would indicate the need for a temporal model. Take particular note of smoking status for men. This was a statistically significant predictor, and was also higher in period 2 for men than in was in period 1.

Reproducibility

Each of the analyses for all aims were performed in `R 4.1.1`.

Code for this analysis is availables from `https://github.com/BIOS6624-UCD/bios6624-JoeFroelicher/project3`, under the branch `project3`. Special note to `@erikaesquinca` for her contributions to the risk over time, and `@ehccooper` for modeling and table making help. These users are cited within the code. Thank you to `@benjamin643` for the use of the `surv_fit()` wrapper to `survfit()`.

Tables and Figures

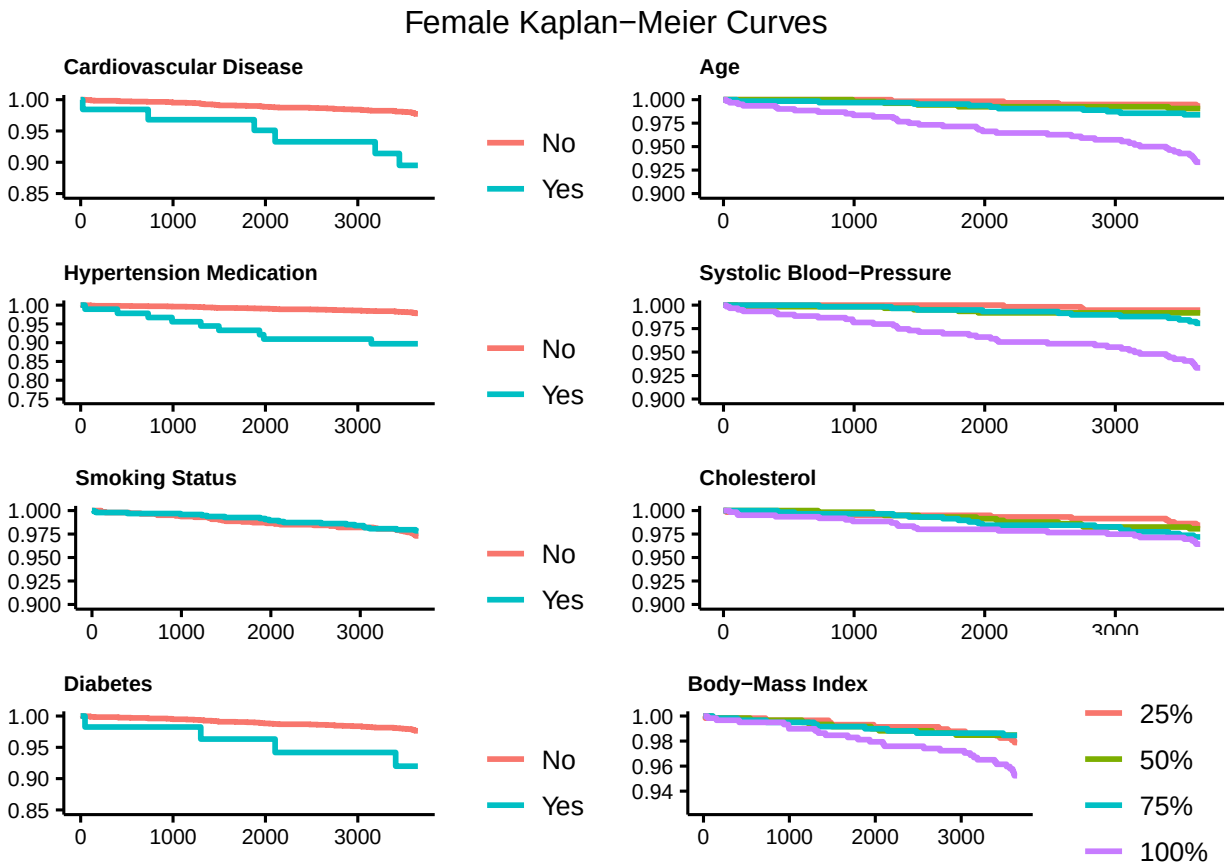

Figure 1. Female Kaplan-Meier Curves for each variable of interest, categorical variables on the left, and continuous on the right, represented in quantile groups, time in days on the independent axis, survival probability on the dependent axis..

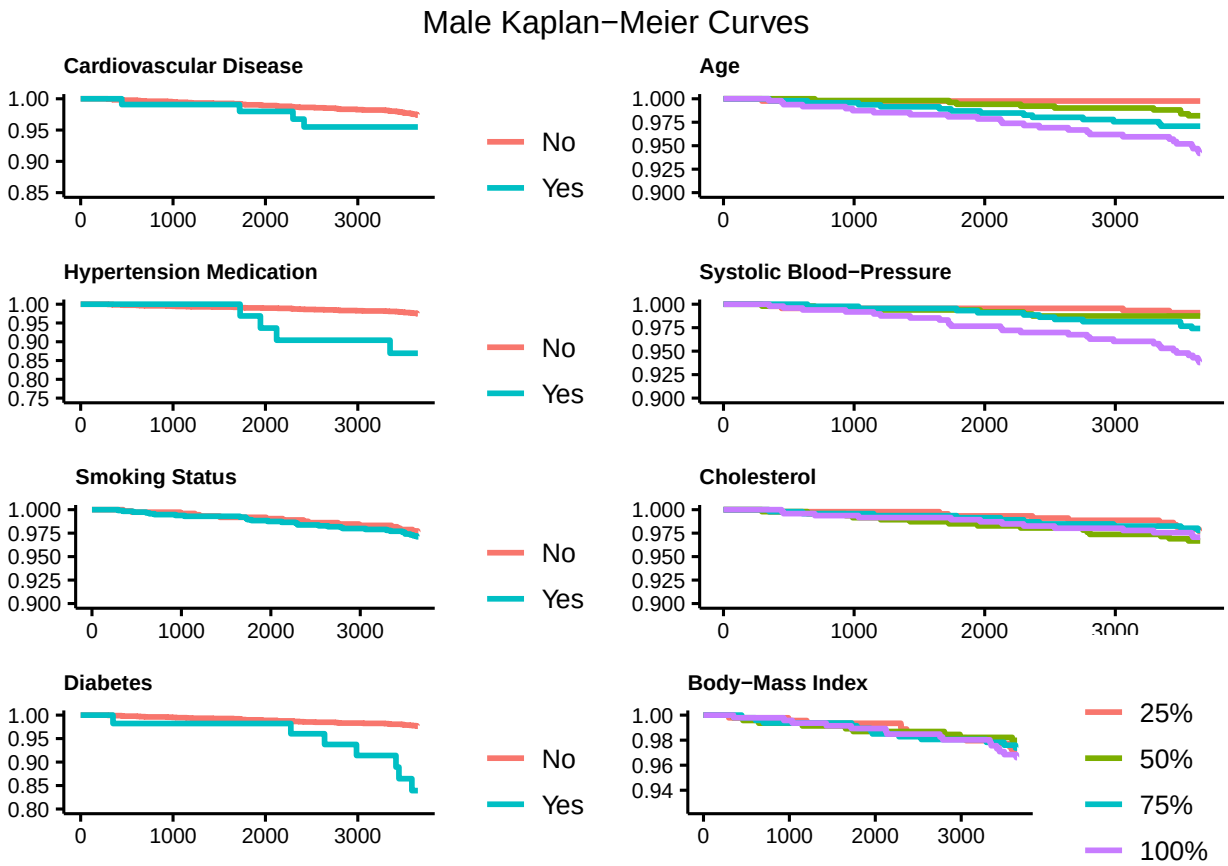

Figure 2. Male Kaplan-Meier Curves for each variable of interest, categorical variables on the left, and continuous on the right, represented in quantile groups, time in days on the independent axis, survival probability on the dependent axis..

		Female		Male	
		Censor	Stroke	Censor	Stroke
n		2321	57	1849	48
Age (mean ± sd)		49.72 ± 8.53	57.23 ± 7.51	49.54 ± 8.69	55.69 ± 7.57
Sys BP (mean ± sd)		132.85 ± 23.56	161.96 ± 34.28	131.03 ± 18.74	151.12 ± 27.89
BP Meds (%)	No	2238 (96.4)	48 (84.2)	1813 (98.1)	44 (91.7)
	Yes	83 (3.6)	9 (15.8)	36 (1.9)	4 (8.3)
Smoker (%)	No	1383 (59.6)	37 (64.9)	733 (39.6)	17 (35.4)
	Yes	938 (40.4)	20 (35.1)	1116 (60.4)	31 (64.6)
Chol (mean ± sd)		239.27 ± 46.08	251.93 ± 45.64	233.63 ± 42.37	237.96 ± 46.19
bmi (mean ± sd)		25.53 ± 4.49	27.20 ± 5.51	26.16 ± 3.37	26.59 ± 4.13
Diabeters (%)	No	2268 (97.7)	53 (93.0)	1800 (97.3)	41 (85.4)
	Yes	53 (2.3)	4 (7.0)	49 (2.7)	7 (14.6)
cvd (%)	No	2263 (97.5)	51 (89.5)	1740 (94.1)	44 (91.7)
	Yes	58 (2.5)	6 (10.5)	109 (5.9)	4 (8.3)
Time to Event (median ± IQR)		3650 ± 0	1982 ± 1876	3650 ± 0	2286.5 ± 2113.75

Table 1. Descriptive Statistics for the Framingham heart data. Continuous variables have means reported, categorical variables have counts and percentages reported.

	Males						Females					
	coef	HR	se(coef)	P-value	lower .95	upper .95	coef	HR	se(coef)	P-value	lower .95	upper .95
Age	0.064	1.066	0.019	0.001	1.027	1.107	0.071	1.074	0.02	<0.001	1.034	1.116
SYS BP	0.033	1.034	0.006	<0.001	1.022	1.046	0.026	1.026	0.004	<0.001	1.018	1.035
Smoke	0.652	1.919	0.316	0.039	1.034	3.564	NA	NA	NA	NA	NA	NA
Diabetes	1.568	4.799	0.423	<0.001	2.096	10.989	NA	NA	NA	NA	NA	NA

Table 2. Results of Cox proportional hazards backwards stepwise model selection for males and females, with raw and exponentiated coefficients. NA values represent risk factors not significant in the Cox P-H model for females.

	Males							Females						
	55	60	65	70	75	80	85	55	60	65	70	75	80	85
Average male/female	0.023	0.031	0.043	0.058	0.080	0.109	0.147	0.020	0.028	0.041	0.058	0.082	0.116	0.162
smoking	0.030	0.041	0.056	0.076	0.104	0.141	0.190	0.025	0.035	0.050	0.072	0.102	0.143	0.199
treated hypertension	0.045	0.062	0.084	0.114	0.155	0.208	0.275	0.029	0.042	0.060	0.085	0.120	0.168	0.232
untreated hypertension	0.022	0.031	0.042	0.058	0.079	0.107	0.145	0.020	0.028	0.040	0.057	0.081	0.114	0.160
diabetes	0.10	0.14	0.19	0.25	0.33	0.43	0.54	0.035	0.050	0.071	0.101	0.142	0.198	0.272
cvd	0.024	0.033	0.045	0.062	0.085	0.115	0.156	0.032	0.046	0.065	0.093	0.131	0.182	0.251
smoking and diabetes	0.14	0.18	0.24	0.32	0.41	0.52	0.64	0.044	0.062	0.088	0.125	0.174	0.241	0.327
smoking and cvd	0.031	0.043	0.059	0.081	0.110	0.149	0.200	0.040	0.057	0.081	0.115	0.160	0.222	0.304
smk and trt hyp	0.059	0.080	0.109	0.148	0.199	0.264	0.346	0.037	0.052	0.074	0.105	0.147	0.205	0.281
smk and unt hyp	0.029	0.040	0.055	0.075	0.103	0.139	0.187	0.024	0.035	0.050	0.071	0.100	0.141	0.196
smk, unt hyp, dbts, cvd	0.14	0.19	0.25	0.33	0.43	0.54	0.65	0.069	0.098	0.138	0.193	0.265	0.358	0.471

Table 3. 10 year risk profiles for various risk profiles of interest, based on Cox proportional hazards model, including different ages for males and females.

		Females		Males	
	level	Period 1	Period 2	Period 1	Period 2
n		2490	2239	1944	1691
age		50.03 ± 8.64	55.66 ± 8.56	49.79 ± 8.72	55.10 ± 8.51
sysbp		133.82 ± 24.46	138.06 ± 24.30	131.74 ± 19.44	135.48 ± 19.90
bpmeds (%)	0	2349 (95.8)	1920 (87.7)	1880 (97.8)	1553 (93.9)
	1	102 (4.2)	270 (12.3)	42 (2.2)	101 (6.1)
cursmoke (%)	0	1484 (59.6)	1392 (62.2)	769 (39.6)	811 (48.0)
	1	1006 (40.4)	847 (37.8)	1175 (60.4)	880 (52.0)
totchol		239.68 ± 46.22	255.67 ± 47.53	233.58 ± 42.36	241.82 ± 42.14
bmi		25.59 ± 4.56	25.65 ± 4.58	26.17 ± 3.41	26.23 ± 3.40
diabetes (%)	0	2428 (97.5)	2158 (96.4)	1885 (97.0)	1617 (95.6)
	1	62 (2.5)	81 (3.6)	59 (3.0)	74 (4.4)
cvd (%)	0	1988 (79.8)	1772 (79.1)	1231 (63.3)	1094 (64.7)
	1	502 (20.2)	467 (20.9)	713 (36.7)	597 (35.3)

Table 4. Change in risk factors over the 10 year risk period (periods 1 and 2).

References

1. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

2. Cox, D. R., & Oakes, D. (2018). Analysis of survival data. Chapman and Hall/CRC.

3. Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69(1), 239-241.