BS852 Final Project
Vanessa Vu

Association of Baseline Systolic Blood Pressure on Mortality Over 22 Years of Follow-Up

Introduction

A subsample of data from the Framingham Heart Study was used to investigate the relationship between systolic blood pressure (SPF) and mortality. The data includes many measures collected at baseline (exam 4) as well as two other variables that are time varying for 2000 total subjects. The main goals of this study are to determine whether SPF at baseline and mortality are associated, if there are any confounders on that association, and if that association differs between males and females.

Methods

To focus on the relationship between SPF at exam 4 (SPF4) as the main predictor and mortality (DTH) as the outcome, time to event must be considered since incidence of death is collected over 22 years of follow up. Thus, a survival analysis is most suitable to determine the association of this relationship, with years after exam 4 at which death was recorded as the time to event variable (SURV). Since there are many measures in the dataset that could potentially act as covariates in the model, a Cox regression model is more ideal than Kaplan-Meier analysis. In the Cox regression model for this study, all observations are assumed to be independent, and all censoring are assumed to be noninformative. The proportional hazards (PH) assumption should also be satisfied for all predictors in the model and all predictors should be uncorrelated to avoid multicollinearity.

To evaluate if the PH and no multicollinearity assumptions are met, the full dataset was explored, except for variables menopause, time to coronary heart diseases (CHD) occurrence, and time to diabetes occurrence. Menopause was removed from further analysis to achieve a pooled analysis with both sexes. The other two variables (CHD_Surv and T2D_Surv) were removed as well because the time varying variables CHD and diabetes (T2D) were addressed in a different manner, so these variables were not needed.

The rest of the variables (SEX, AGE4, CHOL4, CIGS4, SMOKE, SPF4, WGT4, FVC4, BMI4, HTN4, DTH, SURV, CHD, and T2D) were used in the exploratory analysis. Univariate analysis using Cox regression was performed on each variable with the outcome to explore associations of all possible relationships. Correlation analysis on all potential predictors was performed to address potential multicollinearity and ensure the uncorrelated predictors assumption is met. Correlation methods included Pearson's correlation for pairs of continuous variables, chi-square tests for pairs of binary variables, and point biserial correlation for pairs of a continuous and binary variable. Since adding time varying exposures into the model is complex and more difficult to interpret, time varying variables CHD and T2D were addressed by making new variables (CHD4 and T2D4) that are time fixed at baseline to see if these variables can be used instead.

Based on the results of these exploratory analyses, the variables were manually selected for model fitting. Before model fitting, observations with any missing values of the selected variables were removed for a complete case analysis. Model building using Cox regression began with all selected variables. Another model was built after removing nonsignificant covariates and the two models were compared to determine the model of best fit. A ZPH test using Schoenfeld residuals was conducted on each model to ensure the PH assumption is met for all predictors. To address one of the study's goals, interaction between SPF4 and SEX was tested to determine if SEX modifies the association between the main predictor and outcome. No other two-way interactions were tested because there is no reason or prior knowledge to do so. Once the final model was decided, each covariate was evaluated for confounding on the association between SPF4 and mortality adjusting for all other covariates. Confounding by all covariates as a group was also evaluated by comparing the crude model and adjusted model.

Results

From the univariate analysis, all variables individually were found to have a significant association with mortality, except for SMOKE and T2D, using a significance level of 0.05. Although HTN4 and CHD were significant, these variables individually failed to pass the PH assumption based on Schoenfeld residuals (p-value of test for HTN4 is 0.0350 and p-value for CHD is 0.0003). Based on Pearson's correlation between continuous variables, two pairs WGT4 with BMI4 and SPF4 with DPF4 were found to have high Pearson's correlation coefficients (0.79088 and 0.77397 respectively). Using chi-square tests between the binary variables, four pairs were found to have significant association using an alpha level of 0.05: SEX and SMOKE, SEX and CHD,

SMOKE and HTN4, and HTN4 and CHD ($p<0.0001$ for every pair). For the point biserial correlation between the continuous and binary predictors, SMOKE was found to be highly correlated with CIGS4 (correlation coefficient 0.78872) and HTN4 was found to be highly correlated with both SPF4 and DPF4 (correlation coefficients 0.73438 and 0.67176 respectively). Another part of the exploratory analyses was to determine if CHD and T2D can be accounted for as time fixed variables at baseline. However, there were no prevalent cases found for T2D at exam 4 and only 2 cases were found for CHD at exam 4.

To select variables to move forward with model fitting, variables were chosen to be eliminated from further analysis based on non-significance found in univariate analysis, failure to satisfy PH assumption from univariate ZPH testing, and/or failure to adhere to no multicollinearity assumption due to high correlation. Firstly, SMOKE was removed because univariate analysis showed no significant association with the outcome, and it was also highly correlated with HTN4 and CIGS4. Both HTN4 and DPF4 were mainly removed because of its high correlation with the main predictor SPF4, but also because of its high correlation with other variables stated previously. For WGT4, this variable was removed because of it is highly correlated with BMI4. Both WGT4 and BMI4 had similar significance with mortality based on the univariate analysis (WGT4 chi-square is 20.96 and BMI chi-square is 20.07), but since BMI4 gives information about height in addition to weight, it was decided for WGT4 to be removed instead of BMI4. Time varying variable T2D was not significant with the outcome based on univariate analysis so it was removed. Lastly, time varying variable CHD was removed because of its high correlation with two other variables and failure to satisfy PH assumption. These time varying variables could not be accounted for by creating time fixed variables at baseline (CHD4 and T2D4) because no prevalent cases or too few cases are not useful information to detect accurate association. Thus, CHD4 and T2D4 were disregarded.

Moving forward with model fitting using only selected variables (SPF4, SEX, AGE4, CHOL4, CIGS4, FVC4, BMI4, DTH, and SURV), observations with missing values of these variables were removed for complete case analysis, reducing the sample size from 2000 to 1847. An intial Cox regression model was fitted using DTH and SURV as the mortality outcome and the rest of the variables as the predictors (AIC 6900.902) and all predictors passed the PH assumption. However, CHOL4 and BMI4 were found to be nonsignficant so another model was fitted without these two variables to determine if this model is a better fit. This model without CHOL4 and BMI4 has a lower AIC (6897.100), indicating better model fit, and higher p-values for each predictor in the ZPH test, suggesting the predictors conform to the PH assumption better and the hazards ratio (HR) is more stable. Thus, the study proceeded with a Cox regression model of mortality as the outcome, SPF4 as the main predictor, and SEX, AGE, CIGS4, and FVC4 as the covariates (Table 1 for summary statistics of final model).

This model was then used to test for interaction between SEX and SPF4 to determine if the association between SPF4 and mortality changes between the two different sexes. The interaction term was found to be nonsignificant using an alpha level of 0.05 ($p=0.8001$) so no interaction term was added to the model (Table 2). Lastly, confounding was evaluated for the covariates as a group and individually on the association between the main predictor and outcome using the 10% rule of thumb. No confounding was found when comparing the HR between the crude and adjusted models for any of the covariates individually and as a group (Table 3).

Discussion & Conclusions

The final model uses Cox regression to model the association between mortality and SPF4 adjusting for sex, age at exam 4, number of cigarettes smoked per day at exam 4, and pulmonary function at exam 4; SPF4 is significantly associated with mortality adjusting for these variables ($p<0.0001$). As SPF4 increases by 1 unit, the hazard of mortality increases by 1.4% adjusting for other covariates (HR: 1.014, 95% CI: (1.010, 1.017)). The covariates as a group and individually of the final model do not confound the association between SPF4 and mortality, so no confounders were discovered based on the final model. No significant interaction was found between sex and SPF4, indicating no significant difference in association of SPF4 and mortality between males and females (Male HR: 1.014, Female HR: 1.013). A limitation of this study to note is that CHD is significantly associated with the outcome based on univariate analysis but was not included in model fitting. CHD cannot be included as a time fixed variable due to lack of prevalent cases and must be accounted for as a time varying variable, especially since it failed the PH assumption. A potential approach is to treat CHD as a time varying covariate where either stratification by CHD or interaction between time and CHD could be used. However, this is much more difficult to interpret so this study only aims to focus on the association between SPF4 and mortality.

Table 1. Summary statistics of all variables used in final model.

Total Observations: 1847				
Variable	**Mean**	**Std Dev**	**Minimum**	**Maximum**
DTH	0.270	0.444	0	1 (=Death)
SURV	17.888	4.682	0	22
SPF4	133.944	24.247	88	290
SEX	1.556	0.497	1 (=Male)	2 (=Female)
AGE4	49.744	8.528	34	69
CIGS4	8.435	11.580	0	43
FVC4	471.873	109.755	82	833

Table 2. Gender-specific hazards estimates of the association between SPF4 and mortality.

Interaction Term SPF4*SEX (p=0.8001)			
	SPF4 HR*	**95% Wald CI**	
At SEX=1 (Males)	1.014	1.009	1.019
At SEX=2 (Females)	1.013	1.009	1.018

*SPF4 Hazard Ratio is referring to every 1 unit increase in SPF4 (i.e. among men, the hazard of mortality increases by 1.4% for every 1 unit increase in SPF4).

Table 3. Crude and adjusted associations between SPF4 and mortality to evaluate confounding.

Variable Tested for Confounding*	SPF4 HR in Crude Model	SPF4 HR in Adjusted Model	Crude/Adjusted
SEX	1.014	1.014	1
AGE4	1.018	1.014	1.004
CIGS4	1.013	1.014	0.999
FVC4	1.015	1.014	1.001
All Covariates Together	1.022 (Base Model)	1.014	1.008

*Variable tested for confounding refers to that covariate being removed from the model (crude) and then comparing SPF4 association to the adjusted model. All Covariates Together refers to the base model with only SPF4 (crude) compared to all covariates in the model (adjusted).

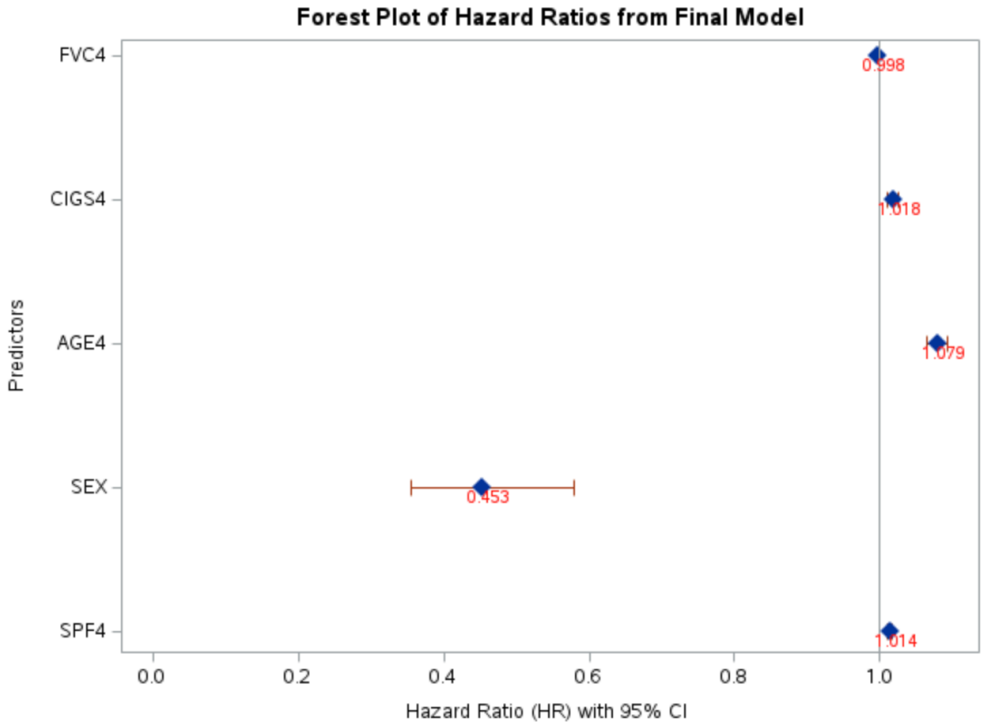

Figure 1. Hazard ratio and 95% CI of each predictor in final model.
*Reference for SEX is 1.

Appendix – SAS Code
(Output is in separate file because it is very long – check *VanessaVu_ProjectSASOutput.pdf*)

```
*/ Vanessa Vu
BS852 Final Project;
/*****************************************/


*/ LOAD IN DATA;
proc import datafile='/home/u63978239/BS852 stats in epi/framdat4.csv'
      out=framdat
      DBMS=csv replace;
      getnames=yes;
      datarow=2;
run;

/*****************************************/


*/ UNIVARIATE ANALYSIS with outcome variable;
proc phreg data=framdat zph;
      model SURV*DTH(0) = SPF4 / rl ties=efron;
run;
*/ SPF significant;
proc phreg data=framdat zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SEX / rl ties=efron;
run;
*/ SEX significant;
proc phreg data=framdat zph;
      model SURV*DTH(0) = AGE4 / rl ties=efron;
run;
*/ AGE significant;
proc phreg data=framdat zph;
      model SURV*DTH(0) = CHOL4 / rl ties=efron;
run;
*/ CHOL4 significant;
proc phreg data=framdat zph;
      class SMOKE(ref='0');
      model SURV*DTH(0) = SMOKE / rl ties=efron;
run;
*/ SMOKE NOT significant;
proc phreg data=framdat zph;
      model SURV*DTH(0) = DPF4 / rl ties=efron;
run;
*/ DPF4 significant;
proc phreg data=framdat zph;
      model SURV*DTH(0) = WGT4 / rl ties=efron;
run;
*/ WGT4 significant;
```

```
proc phreg data=framdat zph;
      model SURV*DTH(0) = FVC4 / rl ties=efron;
run;
*/ FVC4 significant;
proc phreg data=framdat zph;
      model SURV*DTH(0) = BMI4 / rl ties=efron;
run;
*/ BMI4 significant;
proc phreg data=framdat zph;
      model SURV*DTH(0) = HTN4 / rl ties=efron;
run;
*/ HTN4 significant but fails PH assumption;
proc phreg data=framdat zph;
      class CHD(ref='0');
      model SURV*DTH(0) = CHD / rl ties=efron;
run;
*/ CHD significant but fails PH assumption;
proc phreg data=framdat zph;
      class T2D(ref='0');
      model SURV*DTH(0) = T2D / rl ties=efron;
run;
*/ T2D not significant;

/*****************************************/


*/ CORRELATION ANALYSIS;
*/ pearson correlation - continuous covariates and main predictor;
proc corr data=framdat;
var AGE4 CHOL4 CIGS4 WGT4 FVC4 BMI4 DPF4 SPF4;
run;
*/ correlation w chisq - categorical covariates;
proc freq data=framdat;
tables SEX SMOKE HTN4 CHD T2D;
run;
proc freq data=framdat;
tables SEX*SMOKE / chisq;
run;
proc freq data=framdat;
tables SEX*HTN4 / chisq;
run;
proc freq data=framdat;
tables SEX*CHD / chisq;
run;
proc freq data=framdat;
tables SEX*T2D / chisq;
run;
proc freq data=framdat;
tables SMOKE*HTN4 / chisq;
run;
```

```
proc freq data=framdat;
tables SMOKE*CHD / chisq;
run;
proc freq data=framdat;
tables SMOKE*T2D / chisq;
run;
proc freq data=framdat;
tables HTN4*CHD / chisq;
run;
proc freq data=framdat;
tables HTN4*T2D / chisq;
run;
proc freq data=framdat;
tables CHD*T2D / chisq;
run;
*/ point biserial correlation;
proc corr data=framdat;
var AGE4 CHOL4 CIGS4 WGT4 FVC4 BMI4 DPF4 SPF4;
with SEX;
run;
proc corr data=framdat;
var AGE4 CHOL4 CIGS4 WGT4 FVC4 BMI4 DPF4 SPF4;
with SMOKE;
run;
proc corr data=framdat;
var AGE4 CHOL4 CIGS4 WGT4 FVC4 BMI4 DPF4 SPF4;
with HTN4;
run;
proc corr data=framdat;
var AGE4 CHOL4 CIGS4 WGT4 FVC4 BMI4 DPF4 SPF4;
with CHD;
run;
proc corr data=framdat;
var AGE4 CHOL4 CIGS4 WGT4 FVC4 BMI4 DPF4 SPF4;
with T2D;
run;

/*****************************************/

*/ VARIABLES SELECTED: SEX, AGE4, CHOL4, CIGS4, SPF4, FVC4, BMI4, SURV, DTH;

/*****************************************/


*/ MISSING VALUES - complete cases only;
data mainfram; set framdat;
if nmiss(SEX, AGE4, CHOL4, CIGS4, SPF4, FVC4, BMI4, SURV, DTH) = 0;
run;
*/ double check there are no missing values;
proc means data=mainfram n nmiss;
```

```
var SEX AGE4 CHOL4 CIGS4 SPF4 FVC4 BMI4 SURV DTH;
run;
*/ summary of new dataset;
proc summary data=mainfram print;
var SEX AGE4 CHOL4 CIGS4 SPF4 FVC4 BMI4 SURV DTH;
run;

/*****************************************/


*/ INITIAL MODEL;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 CHOL4 CIGS4 FVC4 BMI4 / rl ties=efron;
run;
*/ BMI4 and CHOL4 removed from model because not significant;

*/ FINAL MODEL;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 FVC4 / rl ties=efron;
run;
*/ summary of final model variables;
proc summary data=mainfram print;
var DTH SURV SPF4 SEX AGE4 CIGS4 FVC4;
run;

/*****************************************/


*/ INTERACTION ANALYSIS;
*/ Question 3 - is the association different between males vs females?;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 FVC4 SPF4*SEX / rl ties=efron;
run;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 FVC4 SPF4*SEX / rl ties=efron;
      hazardratio SPF4 / at (SEX='1');
      hazardratio SPF4 / at (SEX='2');
run;
*/ The association does not differ between males vs females;

/*****************************************/


*/ CONFOUNDING ANALYSIS;
*/ base model;
proc phreg data=mainfram zph;
```

```
      model SURV*DTH(0) = SPF4 / rl ties=efron;
run;
*/ fully adjusted model;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 FVC4 / rl ties=efron;
run;
*/ is sex a confounder?;
proc phreg data=mainfram zph;
      model SURV*DTH(0) = SPF4 AGE4 CIGS4 FVC4 / rl ties=efron;
run;
*/ is age a confounder?;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX CIGS4 FVC4 / rl ties=efron;
run;
*/ is cigs a confounder?;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 FVC4 / rl ties=efron;
run;
*/ is fvc a confounder?;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 / rl ties=efron;
run;

/*****************************************/


*/ FOREST PLOT;
proc phreg data=mainfram zph;
      class SEX(ref='1');
      model SURV*DTH(0) = SPF4 SEX AGE4 CIGS4 FVC4 / rl ties=efron;
      ods output parameterestimates=hazards;
run;
title "Forest Plot of Hazard Ratios from Final Model";
proc sgplot data=hazards;
scatter x=HazardRatio y=Parameter /
      xerrorlower=HRLowerCL
      xerrorupper=HRUpperCL
      markerattrs=or(symbol=DiamondFilled size=10);
refline 1 / axis=x;
text x=HazardRatio y=Parameter text=HazardRatio /
position=bottom textattrs=(size=8 color='red');
xaxis label="Hazard Ratio (HR) with 95% CI " min=0;
yaxis label="Predictors";
keylegend / exclude=("Parameter");
run;
```