Exploring Risk Factors Leading to Coronary Heart Disease

Clay Olsen, Grant Hutchings, Sampson Mao

Abstract

Heart disease is a serious condition that affects millions of Americans. This analysis examines demographic, behavioral, and medical risk factors that affect the odds of developing coronary heart disease (CHD). Using a subset of the data collected from the Framingham Heart Study, we develop models to predict CHD outcome. We use forward and backward variable selection methods to find important risk factors. We also fit models using a categorized version of the data, which performs almost as well as the models trained on continuous data. We then consider a model using only demographic and behavioral data. When we use these models to make predictions on unseen data, the performance varied by classification threshold. Additionally, we train models on data balanced with respect to CHD outcome. We use bootstrap to generate these data sets and find very similar results to models trained on the original data set.

KEY WORDS: Logistic Regression, Coronary Heart Disease, Framingham Heart Study, Bootstrap

1. Introduction

Nearly 370,000 Americans die every year from coronary heart disease (CHD). Early detection is crucial in preventing CHD. Early detection allows for treatment to begin quickly, reducing future health complications. In this analysis, we explore medical data collected from a large-scale study in Framingham, Massachusetts, in order to better understand the underlying causes of heart disease. Many risk factors that lead to heart disease, such as high blood pressure and smoking, are known today thanks to the Framingham Heart Study [1]. This study, which began in 1948, is a long-term ongoing cohort study by the National Heart, Lung, and Blood Institute. The first cohort of this study was composed of 5,209 residents from Framingham, Massachusetts. Participants completed surveys and underwent medical screenings every two years. While early data from the Framingham studies provides valuable insights into the relevant risk factors for coronary heart disease, we should be cautious in generalizing results to populations very different from Framingham.

1.1 The Data

The data, available on Kaggle, contains records of the first of three medical examinations of 4,240 of the enrolled residents. The data consists of four categories of predictor variables: demographic, behavioral, medical history, and medical measurements (table 1). The response variable `tenYearCHD` is a column of binary values indicating whether the participant developed heart disease within the ten year period of study. Demographic variables include participant's sex, age, and education level. Behavioral variables include smoking status as well as the number of cigarettes smoked per day. Patient medical history included records of whether patients were taking blood pressure medication, whether they are hypertensive, their diabetic status, and any previous history of a stroke. Finally, doctors took measurements of cholesterol, systolic and diastolic blood pressure, BMI, heart rate, and glucose levels.

It is important to note the difference between blood pressure variables and `prevalentHyp`. Patients who have been diagnosed with hypertension are noted by the `prevalentHyp` variable, but may not have maintained high blood pressure readings during the medical screenings. Additionally, some patients had blood pressure readings at a hypertensive level, but if they were not previously diagnosed for hypertension, they would not be noted under `prevalentHyp`. In short, `prevalentHyp` is associated with a diagnosis, where blood pressure readings and their associated risk categories are from one-time testing.

1.2 Prior Analyses

There have been many research studies done on heart disease using versions of this data set. Since this data set is also available on Kaggle, there are examples of non-researchers who have analyzed this data. One particular user split the data into training and test sets, fit a logistic regression model, and analyzed the accuracy of the model with a confusion matrix [2]. They then refit the model removing insignificant variables, and compare it with the previous model.

Wilson et al. (1998) [4] built models based on the "Offspring Cohort" of the Framingham Heart Study. The researchers decided to use categorical representations of total cholesterol and blood pressure to facilitate ease of use and interpretation of their models by doctors and patients. They found total cholesterol, blood pressure, diabetes, and smoking to be the most important factors contributing to coronary heart disease. Our analysis compares accuracy and predictor choice between models we build, and models using the predictors they found significant, trained on our data.

Table 1: Variables and their Descriptions

	Name	Type	Description
Response	`TenYearCHD`	Binary	Developed CHD within 10 years; 0 - No, 1 - Yes
Demographic	`male`	Binary	0 - Female, 1 - Male
	`age`	Continuous	Age of patient
	`education`	Factor	Education level; 1 - Some high school, 2 - High school or GED, 3 - Some college, 4 - College degree
Behavioral	`currentSmoker`	Binary	Smoker status; 0 - Nonsmoker, 1 - Smoker
	`cigsPerDay`	Continuous	Number of cigarettes smoked per day
Medical History	`BPMeds`	Binary	Prescribed blood pressure medication; 0 - No , 1 - Yes
	`prevalentStroke`	Binary	Previously had stroke; 0 - No, 1 - Yes
	`prevalentHyp`	Binary	Currently hypertensive; 0 - No, 1 - Yes
	`diabetes`	Binary	Currently diabetic; 0 - No, 1 - Yes
Medical Measurements	`totChol`	Continuous	Total Cholesterol level
	`sysBP`	Continuous	Systolic Blood Pressure
	`diaBP`	Continuous	Diastolic Blood Pressure
	`BMI`	Continuous	Body Mass Index
	`heartRate`	Continuous	Heart Rate
	`glucose`	Continuous	Glucose Level

1.3 Questions of Interest

The purpose of this analysis is to deduce which risk factors are the most significant for developing heart disease. To do so, we use various selection methods to fit models. We consider models trained with continuous predictors, and categorical predictors. We are also interesting whether heart disease risks vary between demographics. To answer this question, we fit a model using only demographic and behavioral variables. We hope to build models that are on par with, or superior to those used in prior analyses of this data set. Thus, we include our own categorical models using methods found Wilson et al. (1998). Finally, as early detection of heart disease is conducive to a healthy life, we would like to know how well these models perform out of sample.

2. Exploratory Data Analysis

Since we are predicting a binary value, we use logistic regression to model the data. We first checked for missing values in our data set. 582 out of 4240 rows contained at least one missing value. In these rows, 84 represented patients who had heart disease, and 495 represented those who did not. This corresponds to 14.5% of patients who developed CHD. This proportion is nearly identical to CHD outcomes in the full data set. Given this, we believe dropping rows with missing values will not have a substantial effect on our models.

Next, we explored the relationship between predictor variables and the response with a correlation matrix. We found that many of the predictors in the data set had a low correlation with developing heart disease, with most having a correlation coefficient below 0.1. Diastolic and systolic blood pressure, age, and prevalent hypertension were factors that had higher correlation coefficients. Plotting these predictors against the response confirmed this, as there was only separation in predictors with higher correlation.

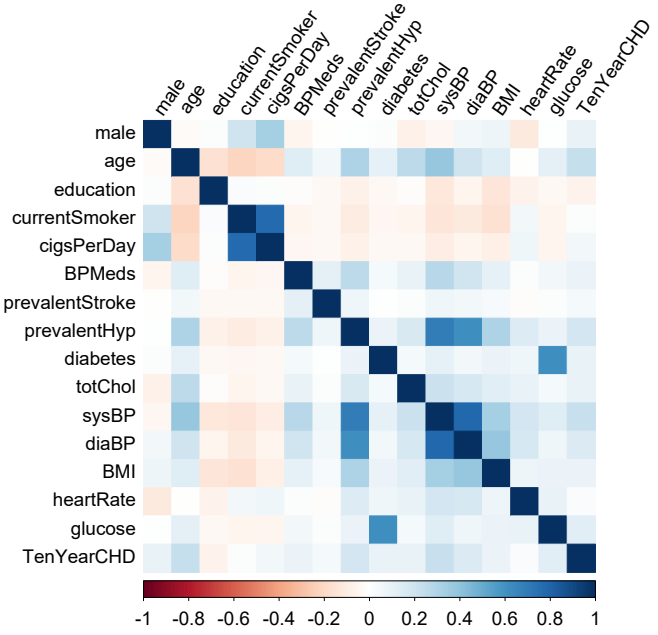

Figure 1: A correlation heatmap of the data. There are a few areas that are much bluer than others, indicating the strong positive correlation between those predictors.

Furthermore, some predictors were correlated with each other, as seen in figure 1. This multicollinearity was mostly due to the redundancy in the recorded data. For example, the glucose values already contain informa-

Table 2: Continuous Models

	Dependent variable:		
	tenYearCHD		
	Forward Stepwise	Backward Stepwise	Significant Subset
age	0.065*** (0.007)	0.066*** (0.007)	0.067*** (0.007)
sysBP	0.016*** (0.002)	0.017*** (0.002)	0.018*** (0.002)
male	0.559*** (0.119)	0.533*** (0.118)	0.525*** (0.117)
cigsPerDay	0.019*** (0.005)	0.019*** (0.005)	0.019*** (0.005)
glucose	0.006*** (0.002)	0.006*** (0.002)	0.006*** (0.002)
BPmeds	0.457* (0.265)	0.481* (0.264)	
totChol	0.002 (0.001)		
Constant	−8.791*** (0.539)	−8.417*** (0.475)	−8.564*** (0.468)
Observations	2,927	2,927	2,927
Log Likelihood	−1,110.272	−1,111.434	−1,113.036
Akaike Inf. Crit.	2,236.545	2,236.868	2,238.072
Note:			*p<0.1; **p<0.05; ***p<0.01

tion about diabetes. For these variables, we kept one and removed the other. For blood pressure, studies show that systolic blood pressure is a more important predictor for heart disease [3]. Therefore, we did not include diastolic blood pressure when building our continuous models.

Looking at the distribution of the predictor variables, we found that only 21 people had a history of stroke. Since this is unlikely to give our model more predictive power, we decided to drop this column.

After getting a better understanding of the data and cleaning the data, we split it into a training set and a testing set with an 80:20 proportion. This split allows us to test the out of sample predictive power of our models.

3. Modeling

3.1 The Continuous Model

From our EDA, we do not expect all the variables from the medical exam to be significant in predicting heart disease. However, based on current knowledge, we expect blood pressure and smoking to be related to heart disease. To find out which predictors are significant in predicting heart disease, we used forward and backward selection. To facilitate this, we used the stepwise variable selection function in R, which adds or removes predictors based on AIC. The results are shown in table 2.

3.1.1 Variable Selection

The models selected by the stepwise forward and backward algorithms had very similar AICs. The initial model before stepwise forward selection is the intercept only model. For the stepwise backward selection, the initial model is the full model. Although the stepwise algorithms produced models with the lowest AICs, they both contain insignificant predictors. For the forward selection model, there are two variables for which, if any one were removed, the resulting model's ability to model `tenyearchd` would not be significantly different. For the backward selection model, there only `BPmeds` is insignificant. While a low AIC model is desired, we do not want to overfit our data. Otherwise, our models may perform poorly on unseen data.

To see if we can reduce the complexity of the model, we examined the analysis of deviance table for the model, and the results are given in table 3. The residual deviance decreases as more predictors are added, which indicates an improved fit. However, starting with the addition of `BPmeds`, there is a smaller decerease in residual deviance. Therefore, we removed `BPmeds` and `totChol` from the stepwise models.

Interestingly, the coefficient for `BPmeds` is large compared to the other variables. Adding this variable at different stages of the model does result in a decrease in residual deviance that is significant. However, based on the z statistic, we never find that this coefficient is significant.

Removing the insignificant predictors only increased our AIC at most by 1.527. The AIC ranged from 2242.95 (null model) to 2500.52 (full model). For residual deviance, it was a change of 5.53 out of 2498.52. Thus, the change in our metrics were very small, and we were able to obtain a simple model of our data.

3.1.2 Analysis

Our continuous model has the form:

$$\begin{aligned}\text{logit}(p) =& \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{sysBP} + \beta_3 \texttt{male} + \\ & \beta_4 \texttt{cigsPerDay} + \beta_5 \texttt{glucose}\end{aligned}$$

The βs in this equation are the effects of the corresponding predictor on the log-odds of developing CHD. The predictor names in the model all represent variables. For example, the effect on the `age` variable is β_1. All of the variables here are continuous except for `male`. It is

Table 3: Analysis of Deviance: Forward Stepwise Selection Model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			2926	2498.52	
`age`	1	152.56	2925	2345.96	4.787e-35
`sysBP`	1	52.36	2924	2293.61	4.63e-13
`male`	1	38.21	2923	2255.40	6.358e-10
`cigsPerDay`	1	16.43	2922	2238.97	5.059e-05
`glucose`	1	12.90	2921	2226.07	0.0003283
`BPmeds`	1	3.20	2920	2222.87	0.07348
`totChol`	1	2.32	2919	2220.54	0.1274

Table 4: Heart Disease Predictions (Continuous)

		Actual			
Threshold		0.50		0.35	
		No	Yes	No	Yes
Predicted	No	615	95	586	77
	Yes	5	16	34	34

	1.00	0.50	0.35
TPR	0%	14.4%	35.6%
TNR	100%	99.2%	94.5%
Accuracy	84.8%	86.3%	84.8%

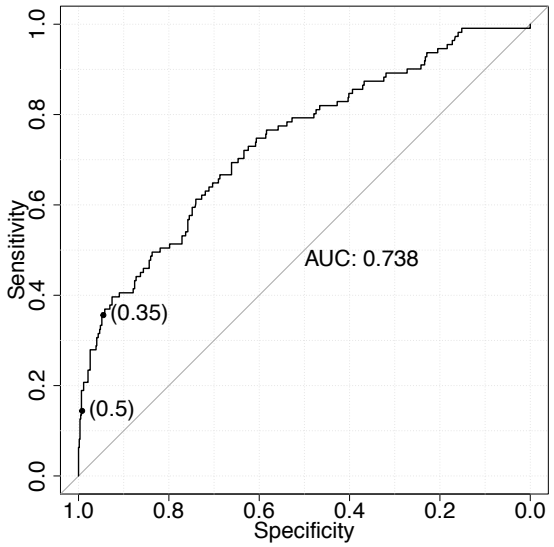

Figure 2: The ROC curve for the continuous model. The points on the plot indicate the location on the ROC curve for the prediction thresholds displayed in parentheses.

clear that the term with β_3 only appears when we consider male patients, otherwise `male` $= 0$. And so, the intercept is interpreted as the log-odds of female respondents of developing heart disease.

Although our model consists of significant predictors, their individual coefficients do appear low, except for `male`. The large range of the continuous variables is likely to be the cause. For example, if we compare a normal systolic blood pressure of 115 to an elevated pressure of 135 (stage 1 hypertension), the person with the pressure of 135 would be $e^{0.018*135-0.018*115} = 1.43$ times more likely to have heart disease. Meanwhile, a smaller difference would have the model show a smaller odds of having heart disease. In comparison, for a binary category, the difference in odds is quite apparent. With a coefficient of 0.525, the change in odds of getting heart disease in males is $e^{0.525} = 1.69$. So if the patient is male, they would be 1.69 times more likely than a female patient to get heart disease if the other factors were the same. It would appear that the effects are low if we compare them individually, but when combined, would give a more pronounced effect. This analysis gives us an understanding of why Wilson et al. (1998) categorized continuous categories like blood pressure and cholesterol, which we discuss in the next section.

3.1.3 Predictions

Now that we have selected a model, we can see how it performs on unseen data. R allows us to output the probabilities directly from log odds. Converting them to binary values with a threshold of 0.5, we obtained a mixture of correctly and incorrectly predicted value, shown in the confusion matrix in table 4. Overall, the model predicted 86.3% of the cases correctly. If our model simply predicted that everybody did not have heart disease, we would have an accuracy of 84.8%. While our model did perform slightly better than the baseline accuracy, the difference is marginal, especially when looking at our true positive rate (TPR).

Our model predicted only 16 true positives out of a possible 111, which is a TPR of 14.4%. This TPR corresponds to a false positive rate (FPR) of 85.6%. Meanwhile, our model predicted 99.2% of the negative cases correctly. With a threshold of 0.5, our model is biased towards negative predictions, being conservative with positive ones. Such a model would most likely be unhelpful in early detection of heart disease, since we are unable to predict the 85.6% of people who actually had it.

A next step would be to lower the probability threshold. With a threshold of 0.35, the number of false negatives decreased, while the number of false positives increased. The number of true negatives also decreased while the number of true positives increased. The TPR is now 35.6%, while the true negative rate (TNR) is 94.5%.

Table 5: Categorized variables. Adapted from Wilson et al. (1998)

(a) Blood Pressure Categories

Systolic BP (mmHg) \ Diastolic BP (mmHg)	<80	80-84	85-89	90-99	>=100
<120	optimal				
120-129		normal			
130-139			high-normal		
140-159				hyp1	
≥ 160					hyp2-4

(b) Cholesterol categories

Total Cholesterol (mg/dl)
<160
160-199
200-239
240-279
>280

As a result, we increased our positive predictive power but sacrificed model sensitivity. Overall, this lowers our model accuracy to 84.8%.

If we compare the TPR and TNR for the two thresholds on figure 2, we can see that their positions on the graph are very close. Despite lowering the threshold by 0.15, we only gained a 21.2 percentage point increase in sensitivity. We would need a large decrease in the probability threshold to raise the TPR, but at the cost of a large drop in TNR and a large increase in FPR, rendering the model ineffective. In order to increase both the TNR and TPR without sacrificing model sensitivity, we need to balance our data and refit our model. We discuss this in section 4.

3.2 Categorical Models

Motivated by Wilson et al. (1998), we fit models using categorized representations of total cholesterol (TC) and blood pressure (BP). Categorizing variables permits the authors to develop score sheets for CHD risk. Point values are assigned to risk categories, and the sum is mapped to a total CHD risk. Total cholesterol and blood pressure are categorized as shown in table 5. Note the maximum risk category between diastolic and systolic blood pressure was used as the patients' overall BP risk category.

The authors claim, models with categorized predictors perform nearly as well as those with continuous. We were able to reproduce this claim, and the results are shown in table 7. We see that AIC is only fractionally reduced when using continuous predictors. Given the similarity in prediction capability, we follow the authors' lead and fit models for males and females separately using categorized TC and BP. We fit four models, summarized in table 6. "Male-Paper" and "Female-Paper" use the predictors specified by the authors while "Male" and "Female" were generated using backward selection based on AIC.

3.2.1 Female Risk Factors

There are many interesting differences between the variables chosen by AIC and the variables used in the paper. Most notably, we find hypertension and education to be significant factors for predicting CHD in women. These predictors were not used in the paper, nor did we find them to be significant for men.

Education is negatively related to CHD risk. Specifically, attaining a bachelor's degree is the number one risk-reducing factor for women, with a **decrease** of 0.828 in log(odds) for developing CHD compared to those without a high school degree. The second most important risk factor for women is hypertension, with a 0.669 **increase** in log(odds). Another interesting discovery is that unlike the other three models, we found that both total cholesterol and blood pressure are insignificant risk factors for women. Recall the difference between `prevalentHyp` and BP risk is described in section 1.1.

Wilson et al. (1998) and the majority of scientific literature on heart disease find total cholesterol to be an extremely important risk factor. We find its absence surprising and possibly reflective only of the population represented in our data.

3.2.2 Male Risk Factors

The differences between our model for males and the model detailed by Wilson et al. are far fewer than those for females. In our models, we use *Glucose Level* and *Cigarettes Per Day* as proxies for *Diabetes* and *Current Smoker* respectively. Considering this, our model chose essentially the same predictors as theirs. We find total cholesterol (TC) and blood pressure (BP) to be the most important predictors. Note that the reference levels for TC and BP–absorbed into the intercept term–are TC between 160 and 199 and BP "normal." The "normal" BP category corresponds to systolic between 120 and 129 and diastolic between 80 and 84.

Our model finds very low total cholesterol (< 160) to be the most important risk factor for CHD. TC below 160 increases log-odds of developing CHD by 1.307 compared to individuals with TC in the range of 160-199. This is surprisingly higher than the increase of 0.784 in log-odds for those with TC greater than 280. The next most important risk factor for males is hypertension. Men with stage 2-4 Hypertension have .827 higher log-odds of developing CHD than those with normal blood pressure. Those with stage one hypertension have a more moderate increase in log-odds of .306. Interestingly, "high-normal"

Table 6: Categorical Models

	Dependent variable:			
	TenYearCHD			
	male	male-paper	female	female-paper
age	0.075*** (0.009)	0.074*** (0.009)	0.072*** (0.011)	0.196 (0.134)
age^2				−0.001 (0.001)
some_HS			−0.329* (0.195)	
HS_or_GED			−0.256 (0.223)	
bachelor			−0.828** (0.412)	
cigsPerDay	0.018*** (0.005)		0.022** (0.009)	
TC_<160	1.307** (0.518)	1.258** (0.520)		−0.880 (1.075)
TC_200-239	0.700*** (0.255)	0.741*** (0.256)		−0.404 (0.267)
TC_240-279	0.709*** (0.263)	0.720*** (0.264)		−0.476* (0.272)
TC_>280	0.784** (0.307)	0.859*** (0.306)		−0.347 (0.285)
prevalentHyp			0.669*** (0.174)	
glucose	0.007*** (0.003)		0.007*** (0.003)	
BP-optimal	−0.258 (0.247)	−0.264 (0.246)		−0.156 (0.287)
BP-high-norm	−0.332 (0.249)	−0.349 (0.249)		−0.055 (0.285)
BP-hyp1	0.306 (0.217)	0.307 (0.217)		0.178 (0.256)
BP-hyp2-4	0.872*** (0.235)	0.876*** (0.235)		0.528** (0.256)
diabetes		1.064*** (0.354)		0.670* (0.366)
currentSmoker		0.565*** (0.162)		0.210 (0.175)
Constant	−6.895*** (0.624)	−6.403*** (0.591)	−6.569*** (0.643)	−9.105*** (3.515)
Observations	1,298	1,298	1,627	1,627
Log Likelihood	−556.704	−555.957	−531.272	−540.199
Akaike Inf. Crit.	1,137.407	1,135.915	1,078.543	1,106.398
AUC	0.7147114	0.6987168	0.7454467	0.7231888
Note:				*p<0.1; **p<0.05; ***p<0.01

blood pressure decreases the odds of CHD. Finally, we see a reduction in log-odds of CHD for those with "optimal" BP compared to those with "normal" BP.

3.2.3 Summary

We find very similar AIC scores between our models and the models proposed by Wilson et al. The difference in AIC between the models for males is negligible, and the difference for females modest with a mere 2.5% reduction compared to their model. Our models have a higher AUC score, but again, only marginally. The most notable differences between our models are the importance of education for women, but not men, and the importance of total cholesterol for men, but not women.

Table 7: Continuous vs. Categorical predictors

	AIC	Accuracy	AUC
Continuous	2201.601	0.836	0.737
Categorical	2220.634	0.84	0.735

3.3 The Demographic Model

The last model we built is one focusing solely on the behavioral and demographic variables. These include age, sex, cigarettes smoked per day, and education level. As in our first model, we used stepwise forward and backward selection to create our demographic model. Both methods resulted in similar models that excluded education. Additionally, we tested for interaction effects, but found none to be significant.

3.3.1 Modeling

Our demographic model has the form:

$$logit(p) = \beta_0 + \beta_1 \texttt{male} + \beta_2 \texttt{age} + \beta_3 \texttt{cigsPerDay},$$

where β_1 is the gender effect multiplied by a binary gender variable `male`, β_2 is the age effect, and β_3 is the effect from cigarettes.

Table 8 summarizes the output of the demographic model. The stars in this table indicate the significance of each coefficient below a 0.01 level. As such, this model explains the response variable better than a reduced model in which one of the variables were to be omitted. It is interesting that education level was not found to be a significant addition to our model. Initially, we hypothesized

those with degrees tend to be from a higher socioeconomic status, which would mean that they have the resources to take better care of their health. In Framingham, this does not seem to be the case as the best demographic model does not include education. At the very least, it seems that age, sex, and cigarette smoking play a significant role in the development of heart disease.

The results of the logistic model show that `age`, `male` and `cigsPerDay` all increased the odds of developing CHD. There is a 0.456 increase in log-odds, given that the patient is male compared to female. Additionally, for every additional cigarette smoked per day there is a 0.018 increase in the log-odds. Furthermore, for every unit increase in age, there is a 0.087 increase in log-odds of developing heart disease. These results are consistent with the other models. The intercept has a coefficient of -6.369, which tells us the baseline log(odds) of a female developing heart disease. For example, if we compare a male patient to a female patient, the male is $e^{0.456} = 1.577$ times more likely to develop heart disease.

Table 8: Demographic Model Summary

	Dependent variable:
	TenYearCHD
age	0.087*** (0.007)
male1	0.456*** (0.115)
cigsPerDay	0.018*** (0.005)
Constant	−6.609*** (0.379)
Observations	2,925
Log Likelihood	−1,129.292
Akaike Inf. Crit.	2,266.584
Note:	*p<0.1; **p<0.05; ***p<0.01

3.3.2 Prediction

We would again like to consider the predictive power of this model to see if we can predict heart disease using only demographic variables. With a probability threshold of 0.5, this model achieved an accuracy of 85.1%. This model performed with accuracy close to that of our other models, but unfortunately did not produce many true positives having a TPR of 0.90%. Since we are trying to predict positive CHD outcomes, we place a high value on true positive rates. Hence, a cutoff of 0.5 is too high for this model.

Like in the continuous model, decreasing the classification threshold results in a model that results in a higher TPR (and thus FPR), while at the same time, a lower TNR. These results are summarized in table 9. Our AUC for this model is 0.694, which dropped by 0.044 compared to the continuous model. An ROC curve is shown in figure 3. It is more difficult to increase TPR for this model, as the curve starts out closer to the diagonal than in our continuous model. So, the threshold would have to be lowered more to get a comparable TPR. We found that the threshold value that optimized the true and false positive rates was approximately 0.1279. Using this cutoff, our model predicts true positives much better at the expense of overall accuracy. While the demographic model did not perform as accurately as the models including the medical data, it still had very comparative predicting power at optimal threshold.

Table 9: Heart Disease Predictions (Demographic)

		Actual			
Threshold		0.5		0.1279	
		No	Yes	No	Yes
Predicted	No	619	110	332	32
	Yes	1	1	288	79

	1.0	0.5	0.1279
TPR	0%	0.90%	71.2%
TNR	100%	99.8%	52.8%
Accuracy	84.8%	84.8%	83.0%

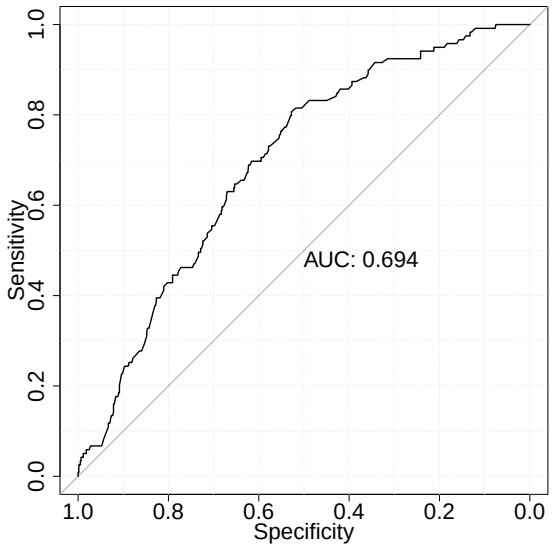

Figure 3: The ROC curve for the demographic model.

4. Bootstrap and Classification Threshold Analysis

This data set is highly unbalanced with respect to CHD outcomes. 84.7% of patients did not develop CHD throughout the study. This means that our models have much more data on negative outcomes. We suspected this could affect the ability of our model to predict positive cases. Figure 4 shows the sensitivity and specificity of the continuous backward selection model over a range of prediction thresholds. We see the cost of high sensitivity is lower specificity.

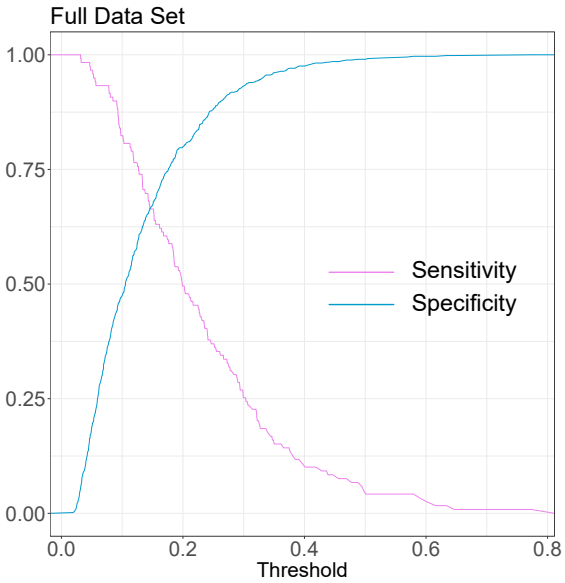

Figure 4: Prediction threshold analysis of sensitivity and specificity for continuous models trained and tested using the full data set.

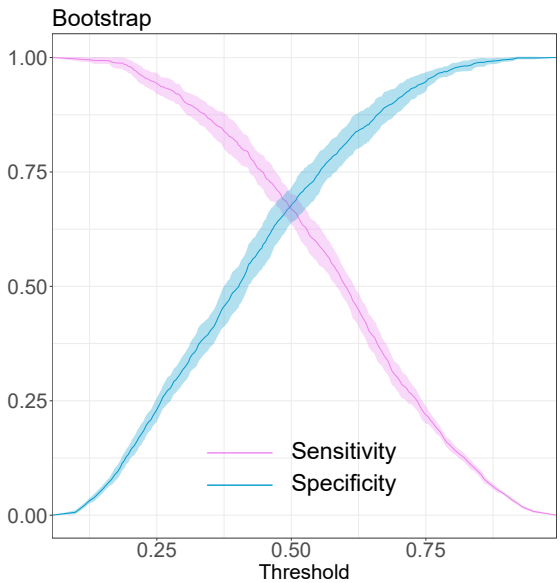

Figure 5: Prediction threshold analysis of sensitivity and specificity for continuous models trained and tested on data balanced with respect to the response variable `tenYearCHD`.

We believe sensitivity to be a vital indicator for the strength of our models. We must be able to correctly predict positive cases of CHD for our model to be useful. We are interested in exploring models trained on a balanced data set. We hope this will give a better trade off between TPR and TNR. We generate 10 data sets using all the true cases, and an equal sampling of false cases without replacement. Data were then randomly split 80/20 into training and testing sets. Ten models were fit to these data sets using a backward selection method. Out of sample sensitivity and specificity were recorded, and averaged over the ten models. Figure 5 shows sensitivity and specificity data over a range of threshold values. Sensitivity and specificity were averaged over 10 models, so a one standard deviation confidence interval is given. We can see that the threshold where the two metrics are equal has shifted considerably to the right. The sensitivity here does not drop as rapidly as it did in the previous figure. However, the positive predictive power did not seem to increase. Table 10 summarizes our findings with bootstrapped data. We choose optimal threshold by maximizing the sum of sensitivity and specificity. At optimal thresholds, bootstrapped models and full data models preform nearly identically. As we value sensitivity, we also consider the situation where we fix a required sensitivity (.75) and maximize specificity under this condition. Again the models are nearly identical. We conclude that bootstrapping does not produce superior models which indicates that a sufficient number of positive cases exist in the original data set to develop our models.

5. Conclusions

Building an accurate prediction model for coronary heart disease is an extremely difficult task. Confounding factors and biological diversity play a major role in health outcomes and make successful predictions of chronic disease challenging. With this in mind, we still find interesting differences between possible risk factors for men and women. We also find differences in the importance of certain risk factors compared with the statistical analysis of Framingham data done by Wilson et al. (1998)

The data show males have a much higher chance of developing coronary heart disease. This is reflected in the large coefficient for males in our continuous and demographic models. For this reason, we agree with Wilson et al. (1998) in their decision to develop prediction models for males and females independently. We also find that age, smoking, and glucose level, or similarly, the presence of diabetes, are significant in all our models. The risk (coefficients) associated with these variables are, however, much lower than the impact of sex. The exception to

Table 10: Continuous Model Performance with Full and Bootstrapped Data

	Full	Bootstrapped
Optimal threshold	0.133	0.468
Sensitivity	0.740	0.720
Specificity	0.623	0.633
Sum	1.362	1.352
Threshold for 75% sensitivity	0.127	0.448
Sensitivity	0.748	0.750
Specificity	0.601	0.592
Sum	1.350	1.342
AUC	0.737	0.733

this is the models based on the set of predictors proposed by Wilson et. al., where we find diabetes and smoking to be far more impactful than in other models.

The most important differences between men and women are education and cholesterol level. Attaining a bachelor's degree is the number one risk-reducing factor for females, while education is not significant for males. On the other hand, total cholesterol is the most important risk factor for men, but our models for women did not find importance.

For our model to be relevant to doctors and patients, we must be able to predict true CHD outcomes. We are willing to accept an increase in false positives for an increase in true positives. Figure 4 show true positive rates fall rapidly with threshold. We believe this may be due to the low proportion of patients who developed CHD. This imbalance in the data means fewer positive cases in the training set by proportion. This results in models that struggle to identify positive cases. We attempted to use bootstrapping as a means to balance our data. When we fit models using the balanced data, we were able to get similar TPR and TNR to our existing models, but without having to lower our threshold considerably. Nevertheless, our TPR did not increase using this method. This was a disappointing outcome, but it gives us confidence that the fit and the predictive power of our original models were not as poor as previously thought.

References

[1] "Framingham Heart Study" (n.d.). Available at `https://www.framinghamheartstudy.org/fhs-about/`.

[2] "Logistic Regression Model On Framingham Dataset" (n.d.). Available at `https://kaggle.com/amanajmera1/logistic-regression-model-on-framingham-dataset`.

[3] Strandberg, T. E., and Pitkala, K. (2003), "What is the most important component of blood pressure: systolic, diastolic or pulse pressure?," *Current Opinion in Nephrology and Hypertension*, **12**, 293–297. `https://doi.org/10.1097/00041552-200305000-00011`.

[4] Wilson, P. W. F., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998), "Prediction of Coronary Heart Disease Using Risk Factor Categories," *Circulation*, **97**, 1837–1847. `https://doi.org/10.1161/01.CIR.97.18.1837`.