Generalized Linear Models for Coronary Heart Disease Prediction

Caleb Ren, Joyce Lu, Abdul Saleh, and Edward Novikov

Stat149 Final Project

INTRODUCTION

One person dies every 37 seconds in the United States from heart disease according to the Center for Disease Control and Prevention[1]. Coronary artery disease is the most common type of heart disease which is caused by the buildup of cholesterol-containing deposits in the arteries that supply blood to the heart. In this project, we use generalized linear models to predict the risk of coronary heart disease given sociodemographic and health risk measures. We use data from an ongoing study on cardiovascular health with participants living in the town of Framingham, MA.

DATA

The study contains 4,238 participants aged $32 - 70$. Of the participants, 43.0% are male; 49.4% are smokers; 27.3% attended college or higher. Each participant was labeled based on their 10-year risk of future coronary heart disease. 15.2% of the participants were labeled at risk, and the remaining 84.8% where labeled not at risk.

Figure 1 shows the distribution of the provided predictors. The predictors are `age`, `sex`, `education level`, `smoking status`, `cigarettes smoked per day`, `BMI`, `heart rate`, `systolic blood pressure`, `diastolic blood pressure`, `blood glucose level`, `total cholesterol level`, `diabetes`, `hypertension`, `previous stroke`, and `blood pressure medication`. Of the sixteen predictors, `total cholesterol level`, `systolic blood pressure`, `BMI`, and `glucose` are skewed right, while `cigarettes smoked per day` appear dramatically zero-inflated.

There are missing values (glucose: 388, education: 105, blood pressure medication: 53, cholesterol: 50, cigarettes: 29, BMI: 19, heart rate: 1). We use the `na.convert.mean` function to deal with missing values. We create a missing value indicator for each of the missing predictors and impute the missing values with the mean for quantitative predictors. We remove the BMI predictor since we find that the observations missing a BMI value have a disproportionately influential and unjustified effect on the model fit.

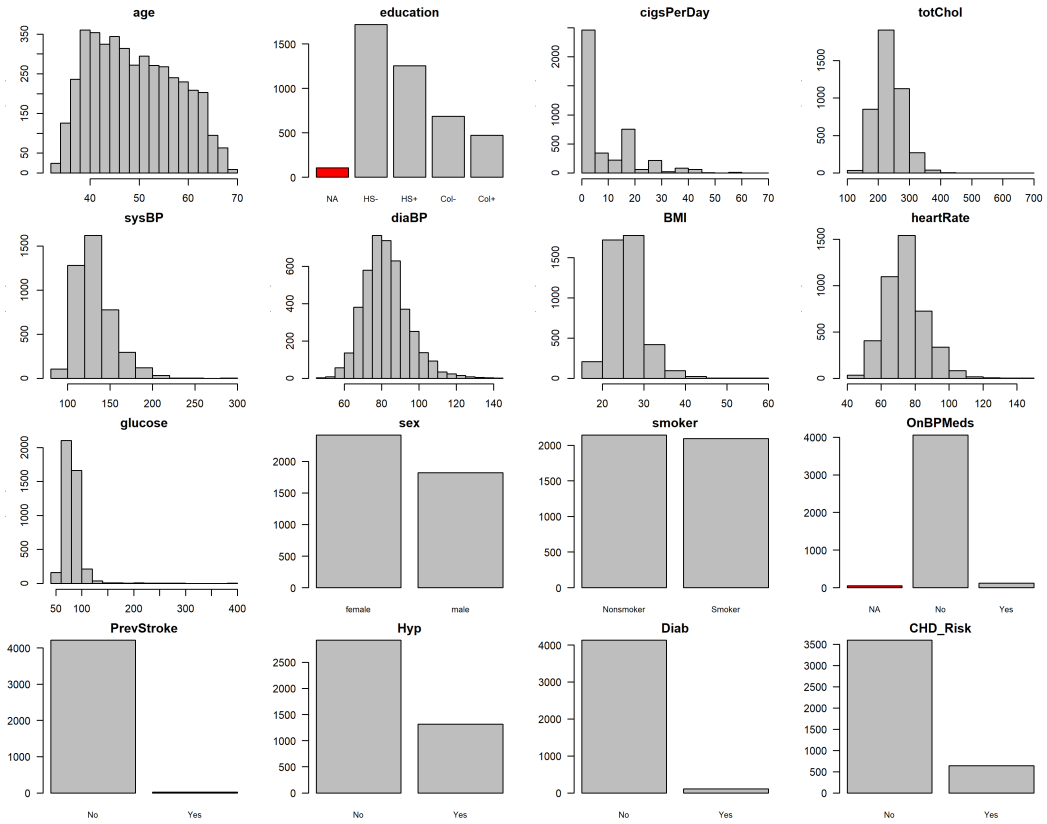


Figure 1. Histograms and bar plots of all the available predictors. NA values are highlighted in red.

After preprocessing the data, we investigated the most telling relationships between `sysBP`, `diaBP`, `age`, and the response. Figure 2 shows the impact of age on future risk. From the data, it appears at a glance that with increasing age, the conditional probability of being considered at-risk for heart disease increases, which makes intuitive sense.

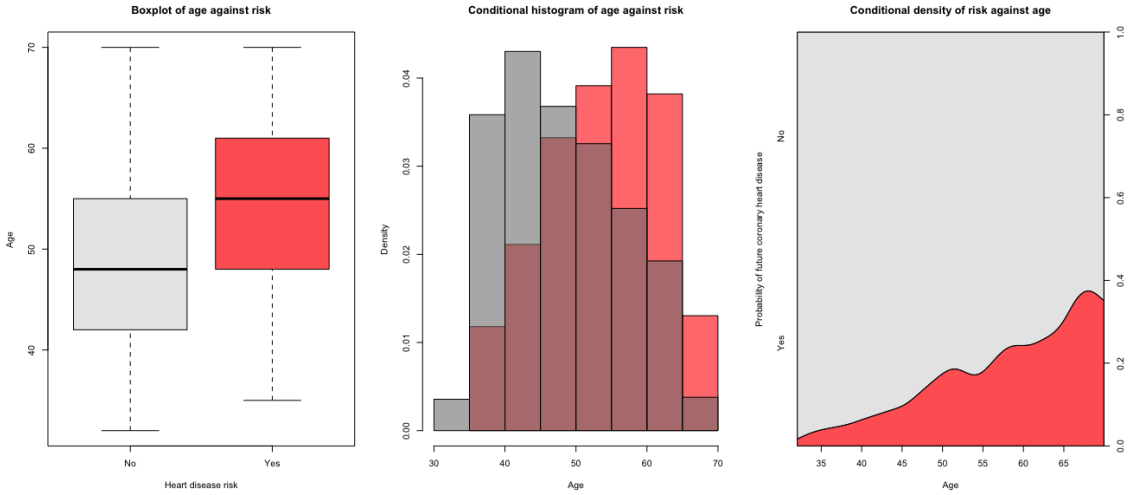


Figure 2. Boxplot of age against coronary heart disease risk, histogram of age conditional on heart disease risk, and conditional density of age against heart disease.

Figure 3 shows the relationship between systolic blood pressure, diastolic blood pressure and risk of coronary heart disease. A strong linear relationship exists between `sysBP` and `diaBP`. From the plot, it does not appear that a particularly strong relationship exists between risk of future coronary heart disease and either blood pressure measures, though there is a slightly greater concentration of at-risk data points toward the higher end of the blood pressure spectrum and a vice versa greater concentration of low-risk data points toward the low end.

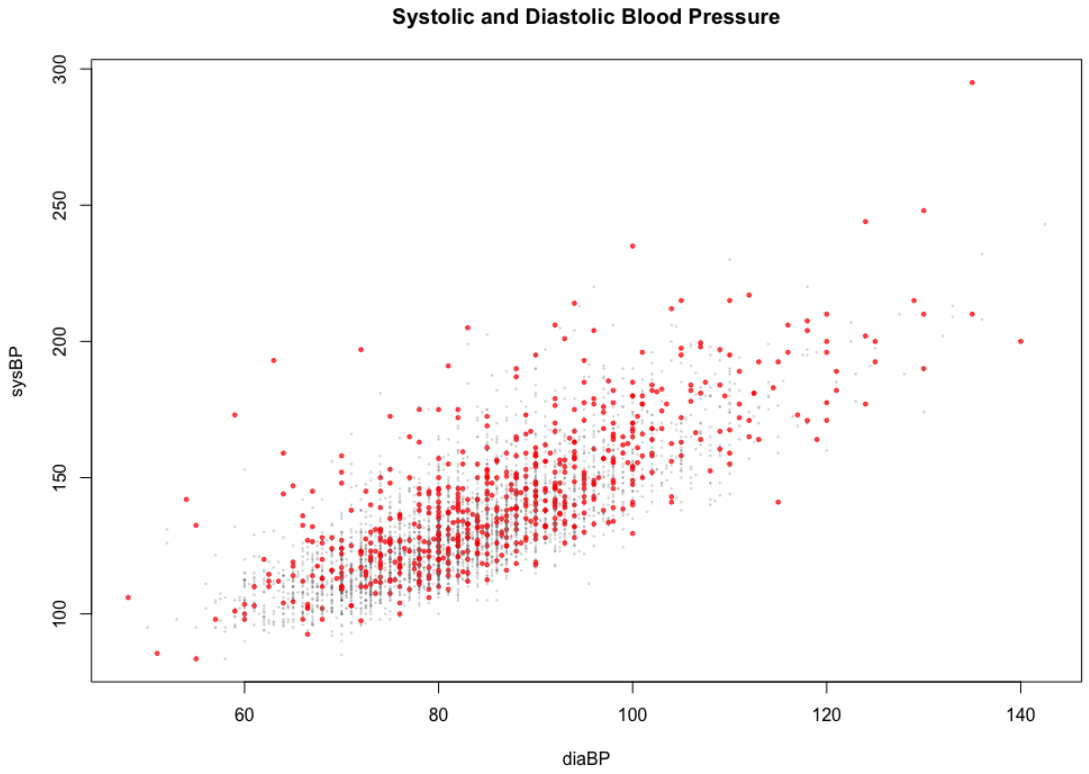


Figure 3. Systolic plotted against diastolic blood pressure, with at-risk data points highlighted in red.

ANALYSIS

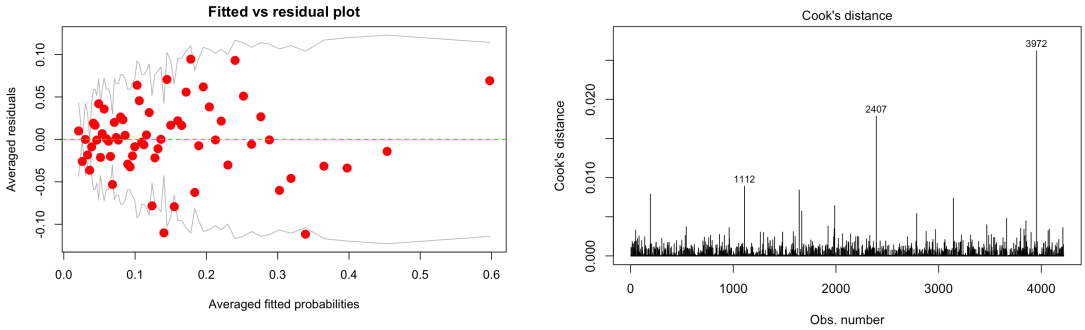
In our analysis, predicting the 10-year risk of future coronary heart disease is treated as a binary classification task. As such, we experiment with generalized linear models, generalized additive models, and decision trees for binary outcomes.

For our first set of experiments, we train logistic, probit, and log-log regression models on the full set of predictors. We find that the different link functions arrive at similar model fits as suggested by their close residual deviances and AIC values (see table 1). The AIC is used as a predictive criterion, where a lower AIC corresponds to better predictive power. We use the logit link for the rest of our analysis for increased interpretability.

Link	Resid. Dev.	AIC
Logit	3188.9	3238.9
Probit	3188.5	3238.5
Log-log	3190.1	3240.1

Table 1. Comparison of different link functions.

Next, we use likelihood ratio tests to find a subset of significant predictors. We start with the intercept only model and add one predictor at a time until no more predictors are deemed significant at the 5% level according to the Chi-squared likelihood ratio test. The resulting seven predictor model contained a significant `missing BMI` predictor, while the `BMI` measurement was deemed not significant. Including the `missing BMI` predictor without considering `BMI` is akin to including a transformed covariate without its original counterpart, resulting in a lack of interpretability. With only nineteen observations having `missing BMI`, we decided to remove this predictor instead of adding the non-significant `BMI` predictor to the model. The final model contains the six predictors: `age`, `sex`, `systolic blood pressure`, `cigarettes smoked per day`, `blood glucose level`, and `previous stroke`. Figure 4 shows diagnostic plots. The residuals look normally distributed and the Cook's distances are small, suggesting a reasonable model fit. We also conduct a low power Hosmer-Lemeshow test for groups of size G = {14, 10, 6, 4} and obtained large p-values indicting no significant lack of model fit. Finally, the variance inflation factors had values less than two, suggesting no multi-collinearity among the predictor variables. The GLM model summary is displayed below.



(a) Fitted vs residuals plot. **(b)** Cook's distance plot.

Figure 4. Diagnostic plots for the fitted GLM.

```
glm(formula = CHD_Risk ~ age + sysBP + cigsPerDay + sex + glucose +
    PrevStroke, family = binomial(link = "logit"), data = chd_imputed)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.462282   0.389647 -21.718  < 2e-16 ***
age           0.064700   0.005926  10.918  < 2e-16 ***
sysBP         0.017070   0.002001   8.531  < 2e-16 ***
cigsPerDay    0.021428   0.003855   5.559 2.71e-08 ***
sexmale       0.486183   0.097204   5.002 5.68e-07 ***
glucose       0.007574   0.001631   4.643 3.44e-06 ***
PrevStrokeYes 1.046918   0.436119   2.401   0.0164 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

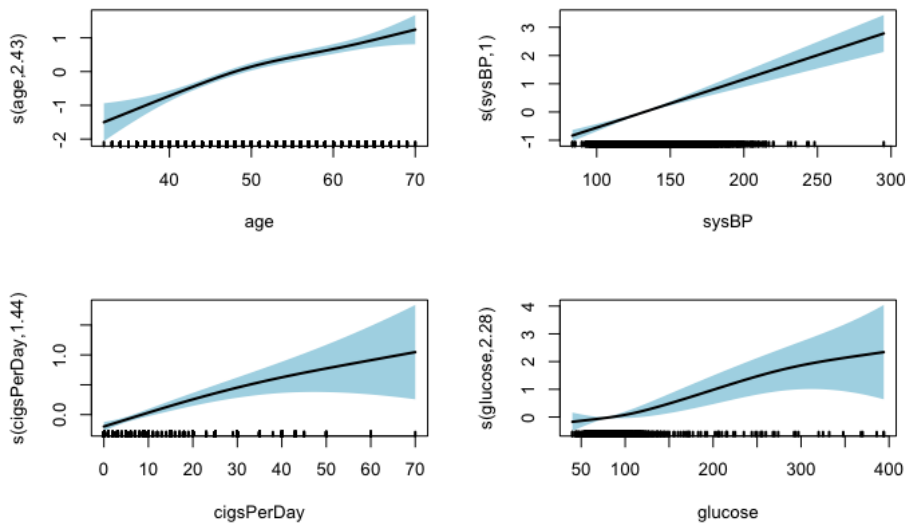


Figure 5. Plots showing the smoothed contributions of the predictors.

```
    Null deviance: 3611.5  on 4237  degrees of freedom
Residual deviance: 3218.6  on 4231  degrees of freedom
AIC: 3232.6
```

We also experiment with generalized additive models. We fit a generalized additive model on the same six predictors from before but estimate the smooth for each of the quantitative predictors: `age`, `cigarettes per day`, `systolic blood pressure` and `blood glucose level`. Figure 5 shows the learned smooths. We can see that `age`, `cigarettes per day`, and `blood glucose level` all have non-linear contributions. This suggests that the generalized additive model provides a better fit than the generalized linear model from before. An analysis of deviance confirms this hypothesis and finds that the GAM leads to a better fit with a p-value $< 2.4\%$. Approximating the estimated degrees of freedom of the smooths, we fit a new GLM with variables `age` and `glucose` transformed. To prevent collinearity with the linear terms, we center the predictors before transforming. Without centering the transformed variables, we find that the linear component of the predictor is not as statistically significant. Incorporating `glucose squared` makes both the linear and quadratic terms non-significant. Including `glucose squared` without `glucose` is significant, but negatively affects the model's residual deviance. Transforming the predictor variables with a cubic does not have a significant contribution. Thus, `age squared` is the only significant transformed predictor when added to the GLM from before. The GAM model summary is shown below.

```
CHD_Risk ~ s(age) + s(sysBP) + s(cigsPerDay) + sex + s(glucose) + PrevStroke

Parametric coefficients:
              Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)   -2.19426    0.07231 -30.345  < 2e-16 ***
sexmale        0.48435    0.09718   4.984 6.23e-07 ***
PrevStrokeYes  1.03826    0.43571   2.383   0.0172 *
---
Approximate significance of smooth terms:
                edf Ref.df Chi.sq  p-value
s(age)        2.427  3.051 118.29  < 2e-16 ***
s(sysBP)      1.001  1.002  73.06  < 2e-16 ***
s(cigsPerDay) 1.442  1.754  28.12 6.46e-07 ***
s(glucose)    2.282  2.858  23.34 4.63e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our final set of experiments used binary classification decision trees. We fit a decision tree on the same six predictors from before and refer to the complexity parameter plot, using the 1-SE rule, to arrive at an appropriate complexity parameter. Figure 6 shows the pruned tree. We use the learned tree to identify useful interactions terms and to provide more interpretable decision boundaries. We can see that age with systolic blood pressure is a potential interaction term, as well as systolic blood pressure with sex and systolic blood pressure with glucose. We conduct likelihood ratio tests to determine significance of the interaction terms. Out of all possible pairwise interaction terms in the tree, with tensor interactions used for quantitative predictors, we find that the interaction of systolic blood pressure with sex is significant with a p-value $< 5.4\%$ when added to the generalized linear model. Three way interaction terms such as age-sysBP-sex and age-sysBP-glucose are also not significant. We also tested the significance of pairwise and three way interaction terms without including age squared in the model, without any improvements. The summary of the final updated GLM is below.

```
glm(formula = CHD_Risk ~ age + age.sq + sysBP + cigsPerDay + sex
    + glucose + PrevStroke + sysBP * sex, family = binomial(link = "logit"),
    data = chd_imputed_transformed)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.3310588  0.4644444 -17.938  < 2e-16 ***
age            0.0725594  0.0071830  10.101  < 2e-16 ***
age.sq        -0.0012739  0.0006724  -1.895   0.0581 .
sysBP          0.0139507  0.0025403   5.492 3.98e-08 ***
cigsPerDay     0.0213494  0.0038720   5.514 3.51e-08 ***
sexmale       -0.5627051  0.5542798  -1.015   0.3100
glucose        0.0075903  0.0016233   4.676 2.93e-06 ***
PrevStrokeYes  1.0501159  0.4351149   2.413   0.0158 *
```

```
sysBP:sexmale  0.0075303  0.0039144   1.924   0.0544 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

    Null deviance: 3611.5  on 4237  degrees of freedom
Residual deviance: 3210.9  on 4229  degrees of freedom
AIC: 3228.9
```

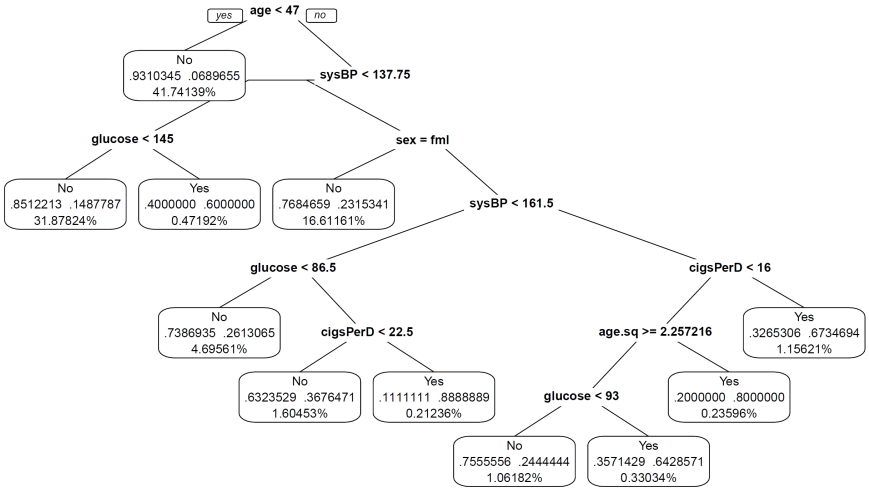


Figure 6. The pruned classification tree.

DISCUSSION

This section interprets the learned models to better understand coronary heart disease and at-risk populations. Consider the person in our sample. The average person is 49.6 years old, is female, has a systolic blood pressure of 132.4, smokes 9.0 cigarettes per day, has a blood glucose level of 82.0, and has never had a previous stroke. According to the final GLM, the predicted 10-year risk coronary heart disease for the average person in our data is 24.7%. The odds of coronary heart disease risk increase by a multiplicative factor of 1.07 with every additional year of age and 1.02 with every additional cigarette. Higher glucose and blood pressure are also associated with an increased risk. Figure 7 shows the predicted risk vs the quantitative predictors with one predictor being varied and all others held constant for the average person. It is important to note that the coefficients found here indicate association and not necessarily causality.

For health economists and medical researchers, it may be important to note that high systolic blood pressure, frequent habitual cigarette smoking, high glucose level, and having had a stroke significantly increase a patient’s 10-year risk for coronary heart disease. While in our final model, sex was no longer a significant predictor, this does not mean necessarily that it should be ignored, or that males do not exhibit coronary heart disease risk factors (such as high systolic blood pressure) more frequently. In addition to keeping close tabs on these risk factors, we would advise investigating potential relationships between these factors and other demographic data (such as income and race) so as

to better target at-risk communities. While certain factors may not end up being statistically significant in a model, they can act as proxies or signals for underlying risk factors; even in our exploratory data analysis, we found positive associations between systolic and diabolic blood pressure, age and hypertension, glucose and diabetes, and hypertension and systolic blood pressure.

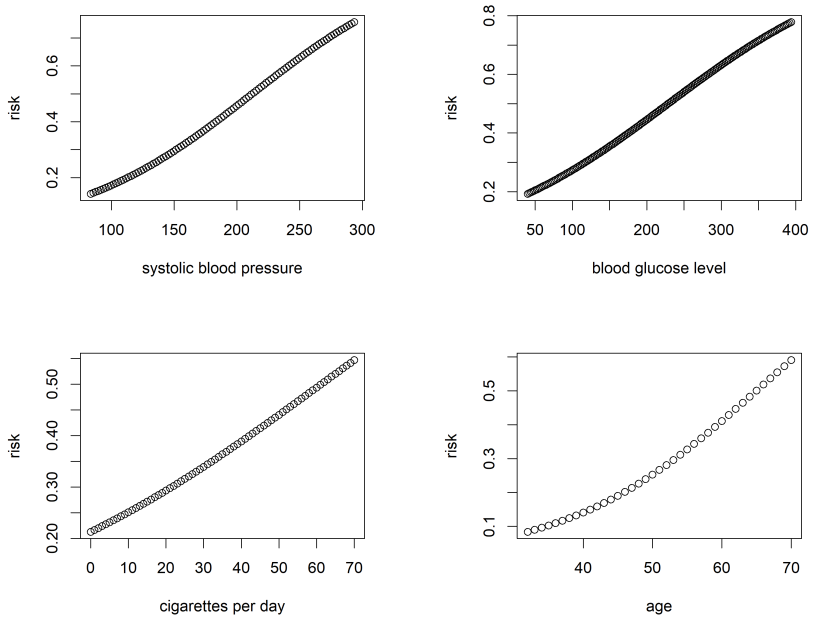


Figure 7. The predicted risk vs the quantitative predictors.

CONCLUSION

To summarize, our final model is an eight-predictor GLM including six linear predictors (`age`, `sysBP`, `cigsPerDay`, `sex`, `glucose`, and `PrevStroke`), one quadratically transformed predictor (`age.sq`), and one interaction term (between `sysBP` and `sex`). The final predictors were selected and refined using results from a GAM model, which indicated to us that age had a nonlinear contribution, and a decision tree, which highlighted the interaction between sex and systolic blood pressure.

Due to the nature of the task, our modeling approach prioritized interpretability and generalizability. In a high-stakes healthcare setting where the goal is to find relationships between coronary heart disease and sociodemographic and health factors, we believe it is important for modeling decisions to be supported by or consistent with existing medical knowledge, rather than rely entirely on inference from one dataset. As a result, we chose to use a modified version of the GLM with a logit link as our final model, despite finding the GAM to be significant in a likelihood ratio test.

Originally, we suspected that log-transforming the right-skewed variables pointed out in the Data section (`totChol`, `sysBP`, `BMI`, `glucose`) might have an effect on the significant predictors, but this turned out not to be the case. In addition, several predictors that we expected to interact in a meaningful way due to results from exploratory data analysis were ultimately not significant in the fi-

nal model. One example of this is the pronounced difference in smoking habits between men and women (with men smoking more frequently). An explanation could be that once other factors were accounted for, these interactions were no longer pronounced.

The main limitation of our model is that the sensitivity is too low. Out of 644 patients in the data set that were identified as having 10-year risk of coronary heart disease, our model only correctly identified 55 of them, for a true positive rate of 8.5%. The specificity of our model was much stronger, with a true negative rate of 99.2%. In an actual healthcare setting, we would prefer high sensitivity to high specificity, as in most cases a false positive is less dangerous than a false negative. Another related limitation of our model is that residuals tended to increase as fitted probabilities increased. If we were to continue the project, we would likely focus on correcting the undersensitivity of the model.

REFERENCES

1. Heart disease facts, Dec 2019. URL `https://www.cdc.gov/heartdisease/facts.htm`.