A Breath of Trouble : Exploring Tobacco's Impact on Coronary Health

Introduction

What link exists between smoking and cardiovascular disease?
According to the 2014 Surgeon General's Report on smoking and health, smoking is a major cause of cardiovascular disease (CVD) and accounts for one in every four CVD-related fatalities. Early CVD symptoms can appear in smokers of less than five cigarettes per day as well. The risk of cardiovascular disease rises with daily cigarette consumption and prolonged smoking.

In what ways does smoking impair cardiovascular health?
The chemicals in cigarette smoke inflame and swell the blood vessel lining cells. This may cause blood vessel narrowing and increase the risk of developing a number of cardiovascular diseases, including coronary heart disease (CHD). It happens when clots or plaque constrict the arteries that supply blood to the heart muscle. Tobacco smoke contains chemicals that cause blood in veins and arteries to thicken and clot. A clot-related blockage may cause a heart attack and unexpected death. [3]

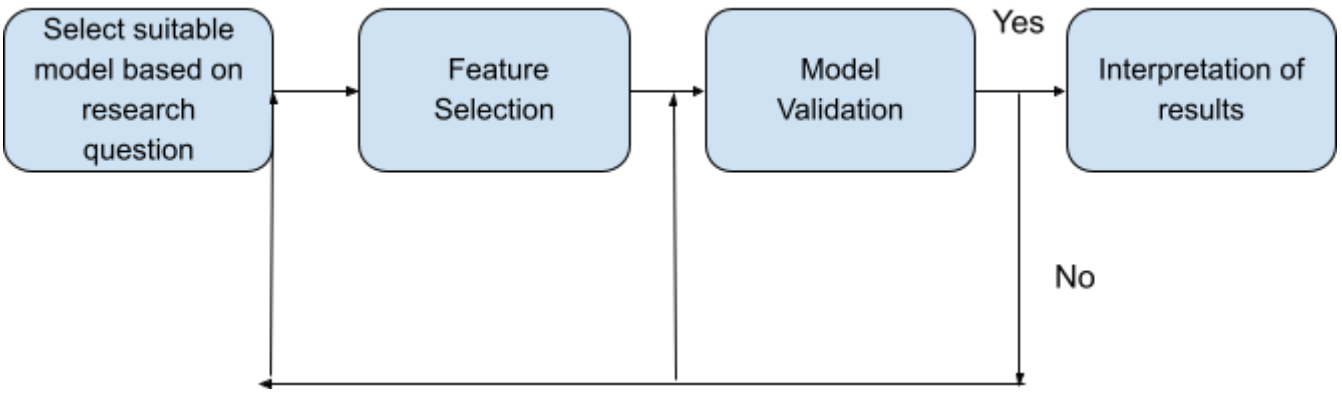
Research question:
As tobacco is mostly used for smoking due to its high nicotine content, along with other risk factors that may be possible causes of CHD, we conduct a statistical exercise to evaluate if tobacco usage has any impact on the diagnosis of CHD. We select "Replication Data for: South African Heart Disease" as a dataset from the HARVARD Dataverse.[2]

Methods

Generalised linear models: We have an outcome variable which is not continuous or normally distributed.

Distribution of Y (outcome variable) : Binomial i.e. binary data.
Canonical Link function: Logistic: $g(p) = log(p/1 - p)$

Strategy:



Metadata

Dataset has 10 Variables and 462 Observations.

1. Class : This is the target variable indicating the presence of coronary heart disease. A value of -1 suggests the absence (negative) of CHD, whereas a value of 1 indicates the presence (positive) of CHD. We then transform -1 to 0, which indicates absence of CHD.

2. Systolic Blood Pressure (sbp): This measures the maximum blood pressure during contraction of the ventricles.

3. Cumulative Tobacco (tobacco): This quantifies the cumulative amount of tobacco consumed, measured in kg.

4. Low-Density Lipoprotein Cholesterol (ldl): Often referred to as bad cholesterol, higher levels of LDL can lead to a build-up of cholesterol in arteries, potentially increasing the risk for CHD.

5. Adiposity: Measure of body fat .

6. Family History of Heart Disease (famhist): This variable indicates whether the individual has a family history of heart disease (1 for present, 0 for absent).

7. Type-A Behavior (typea): Characterised by competitiveness, impatience, and a sense of urgency, this behaviour pattern has been studied for its potential link to heart disease.

8. Obesity: Body mass index (BMI).

9. Alcohol: This measures the amount of alcohol the individual currently consumes.

10. Age: This refers to the age at which the individual was diagnosed with CHD

Exploratory Data Analysis

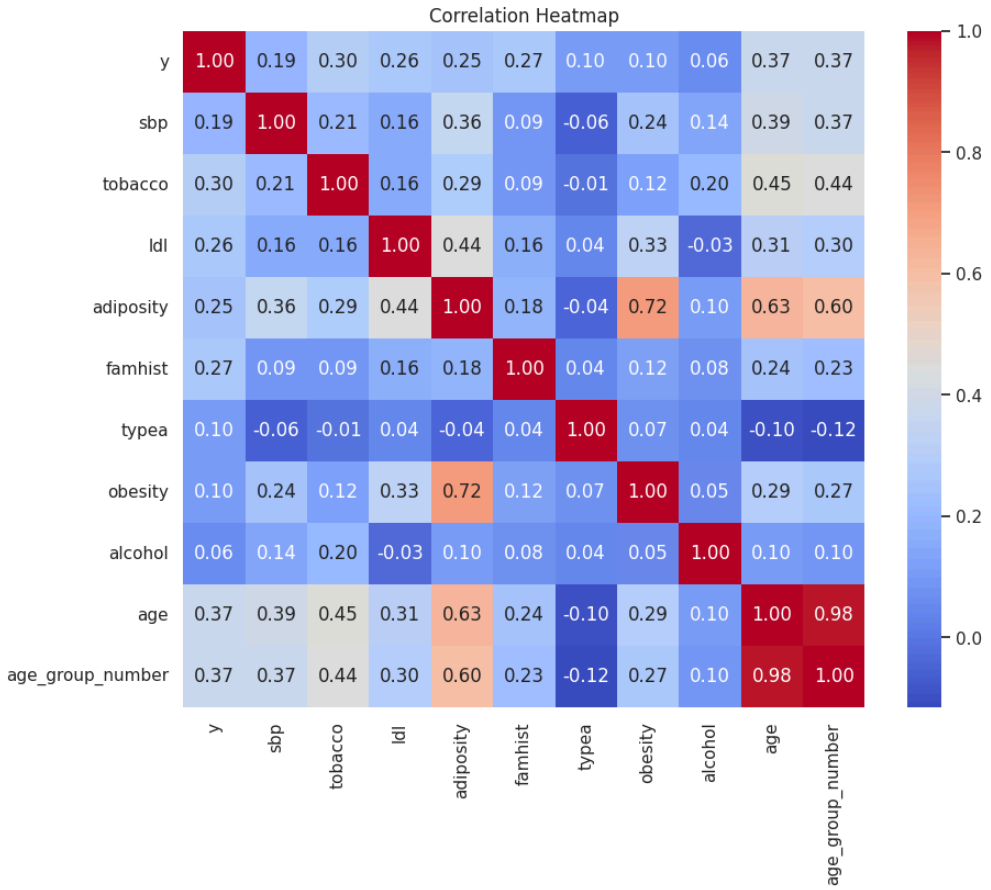


Figure 1. Correlation Heatmap

adiposity and obesity show a high positive correlation (0.72), which suggests that as adiposity increases, obesity tends to increase as well.

ldl and adiposity (0.44), indicating that higher levels of ldl cholesterol might be moderately associated with greater adiposity.

y (CHD) shows moderate correlations with age (0.37), suggesting that age may be a moderately strong predictor of CHD.

y (CHD) shows moderate correlations with tobacco (0.30), suggesting that it may be a moderately strong predictor of CHD. And a correlation of 0.27 , 0.25 with family history (famhist) and adiposity respectively indicating significant association.

CHD with factors like sbp (systolic blood pressure), alcohol, typea (type A behaviour) and obesity show weak positive correlations, suggesting that these may or may not be associated with CHD.

No negative correlations observed in this heatmap.

Causal links between variables

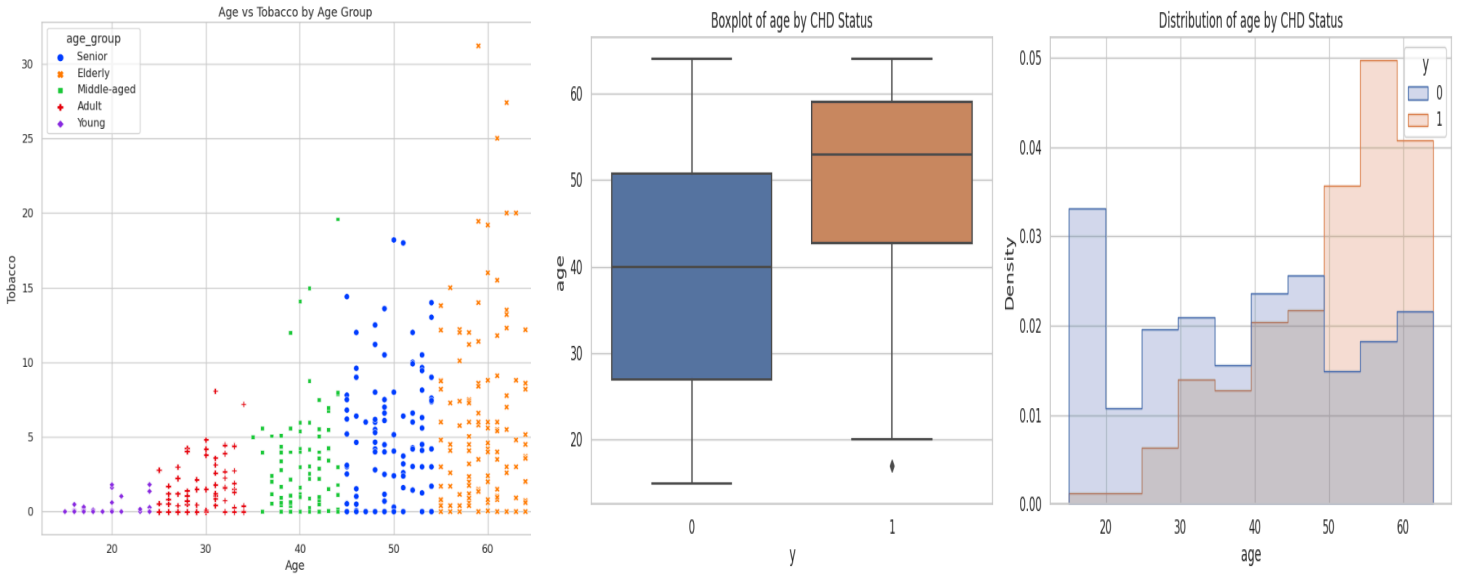


Figure 2. a) Age Vs Tobacco (in Kg) Scatter plot. b) Boxplot of Age by CHD status
c) Distribution of age by CHD status

CHECKING FOR POTENTIAL CONFOUNDERS

The change in the tobacco coefficient after considering age groups as a factor suggests that age does, in fact, have a confounding effect on the connection between tobacco use and CHD. Age impacts the risk of CHD and is likely connected with tobacco use Thus it is a confounder.

Furthermore, the fall in AIC from 558.65 to 526.71 when age is included indicates that the model with age as a factor fits the data better.

CHECKING FOR POTENTIAL MODERATORS

Family History could serve as a moderator. Those with a family history of CHD may have a higher risk of developing CHD due to varying ldl levels.
Thus, we introduce the interaction term, ldl* famhist_f.

After inserting an interaction term between LDL and family history (ldl*famhist_f), the model reveals that family history moderates the influence of LDL cholesterol levels on the risk of CHD. This interaction term is significant, showing that the connection between LDL levels and CHD risk varies between those with and without a family history of the condition.
The positive coefficient for the interaction term shows that for individuals with a family history of CHD, higher LDL levels may be related with a bigger increase in CHD. [1]

Model Selection

In our model selection procedure, we started with a complete model that included all of the predictors. A further comparison with a simplified model using the Likelihood Ratio Test (LRT) showed no significant loss in performance after the elimination of multiple predictors, implying that a simpler model was equally good. Further LRTs revealed that we can rule out the 'obesity' predictor. Finally, including an interaction term between 'ldl' and 'famhist_f' significantly improved the model's fit. This step-by-step technique to balancing model complexity and fit highlights the final model's usefulness.

FINAL MODEL : y ~ tobacco + ldl + famhist_f + typea + age_f + ldl:famhist_f

Cross validation

In the model validation step, cross-validation was used to evaluate the performance of two logistic regression models. The first model, which used a specific set of predictors such as tobacco use, LDL cholesterol, family history, Type A behaviour, age, and an interaction between LDL and family history, achieved a cross-validated accuracy of 73.15% and a Kappa of 0.372.

The second, more sophisticated model, which included all 13 available predictors, had a slightly lower accuracy of 71.41% and a Kappa value of 0.338. These findings highlight the effectiveness of the parsimonious model, which, by using fewer predictors, not only simplified the analysis but also produced higher predictive accuracy.

Results

Table 1: Bayesian Logistic Regression Model Results

Coefficient	Estimate	Std. Error	95% CI		Rhat	Bulk_ESS	Tail_ESS
			Lower	Upper			
Intercept	-4.51	0.82	-6.13	-2.94	1.00	3131	2370
tobacco	0.11	0.03	0.05	0.16	1.00	3274	2043
ldl	0.05	0.07	-0.10	0.19	1.00	2316	2235
famhist_f1	-0.58	0.52	-1.62	0.44	1.00	1892	2161
typea	0.03	0.01	0.01	0.06	1.00	4474	2118
age_f2	0.43	0.45	-0.47	1.31	1.00	1516	1886
age_f3	0.78	0.43	-0.07	1.65	1.00	1571	1813
age_f4	0.72	0.42	-0.11	1.53	1.00	1592	1609
age_f5	1.57	0.41	0.77	2.38	1.00	1560	1754
ldl:famhist_f1	0.32	0.10	0.12	0.52	1.00	1761	1959

y ~ tobacco + ldl + famhist_f + typea + age_f + ldl*famhist_f

Intercept (0.01099846): This figure shows the probability of the result being 1 when all predictor variables are zero. Given its close proximity to 0, it shows very low probability of the outcome being '1' in the absence of any predictors.

tobacco (1.11627807): Assuming all other variables remain constant, each one-unit increase in tobacco usage increases the odds of the outcome being 1 by approximately 11.6%.

ldl (1.05127110): A one-unit rise in LDL cholesterol levels raises the odds of the outcome being 1 by about 5.1%, holding all other variables constant.

famhist_f1 (0.55989837): Individuals having a family history (famhist_f1 = 1) are approximately 44% less likely to have the outcome '1' than those without a family history, all other factors remaining constant. This finding is paradoxical. The combination of family history with ldl cholesterol results in a more relevant observation.

typea (1.03045453): For each one-unit rise in the Type A behaviour score, the odds of the outcome being 1 increase by around 3%, provided other factors remain constant.

age_f2 (1.53725752): Being in age group 2 rather than the reference age group raises the odds of the outcome being 1 by about 53.7%.

age_f3 (2.18147227): Being in age group 3 raises the odds of the result being 1 by approximately 118% when compared to the reference group.

age_f4 (2.05443321): Being in age group 4 raises the odds of the result being 1 by about 105% when compared to the reference group.

age_f5 (4.80664819): Being in age group 5 raises the odds of getting 1 by approximately 381% when compared to the reference age group.

ldl_famhist_f1 (1.37712776): The interaction term indicates that for individuals with a family history of condition famhist_f1, the effect of each one-unit increase in LDL on the odds of the outcome being 1 is further increased by approximately 37.7%, beyond the effect of LDL alone.

Inference

More usage of tobacco can easily increase the patient’s risk of being susceptible to having CHD
.
Being in the Senior age group (55 to 64) has the highest risk of CHD followed by the age group (34 to 44). So they have to be more cautious with their heart health.

Patients whose parents have had CHD should make sure to keep their ldl levels low and be more careful with their heart health.

Patients displaying aggressive behaviour do not have any significant effect on their CHD status.

Discussion

The model can be improved in the following ways:

1. Gathering more data. Training the model using a larger dataset can improve the accuracy and predictive power.
2. The algorithm used is logistic regression. We can choose more complex algorithms to classify the outcome variable such as Random forest, Decision trees and SVMs.
3. Researching more domain knowledge from subject matter experts can further help in choosing a prior for the model and variables that might have more impact on the outcome variable.

References

1 "Family History of Coronary Heart Disease? It Might Be Your Genetics" Healthline. [Online]. Available: https://www.healthline.com/health/is-coronary-artery-disease-genetic. [Accessed: 21 Mar. 2024]

2 C. Bartley, "Replication Data for: South African Heart Disease," Harvard Dataverse, 2016. [Online]. Available: https://doi.org/10.7910/DVN/76SIQD. [Accessed: 20 Mar. 2024]

3 Centers for Disease Control and Prevention, "The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General," U.S. Department of Health and Human Services, [PDF]. Available: https://www.cdc.gov/tobacco/sgr/50th-anniversary/pdfs/fs_smoking_cvd_508.pdf. [Accessed: 21 Mar 2024].

4 ApokalypsePartyTeam, "Simple examples to understand what confounders, colliders, mediators, and moderators are, and how to control for variables in R with regression and propensity score matching," R-bloggers, 01-Jan-2022. [Online]. Available: https://www.r-bloggers.com/2022/01/simple-examples-to-understand-what-confounders-colliders-mediators-and-moderators-are-and-how-to-control-for-variables-in-r-with-regression-and-propensity-score-matching/. [Accessed: 22-Mar-2024].

Supplementary Information

Confounder:

Table 1: Model 1: Logistic Regression Model with Tobacco Predicting Coronary Heart Disease

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.18943	0.13900	-8.557	< 2e-16 ***
tobacco	0.14527	0.02476	5.866	4.46e-09 ***

Note: Null deviance: 596.11 on 461 degrees of freedom. Residual deviance: 554.65 on 460 degrees of freedom. AIC: 558.65. Number of Fisher Scoring iterations: 4.

Table 2: Model 2: Logistic Regression Model with Tobacco and Age Group Predicting Coronary Heart Disease

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.39450	0.71901	-4.721	2.35e-06 ***
tobacco	0.08389	0.02578	3.254	0.00114 **
as.factor(age_group_number)2	1.85820	0.77580	2.395	0.01661 *
as.factor(age_group_number)3	2.42794	0.76096	3.191	0.00142 **
as.factor(age_group_number)4	2.47514	0.75920	3.260	0.00111 **
as.factor(age_group_number)5	3.13206	0.75114	4.170	3.05e-05 ***

Note: Null deviance: 596.11 on 461 degrees of freedom. Residual deviance: 514.71 on 456 degrees of freedom. AIC: 526.71. Number of Fisher Scoring iterations: 6.

The change in the coefficient for tobacco after including age groups as a factor indicates that age does indeed have a confounding effect on the relationship between tobacco use and CHD. Age affects the risk of CHD, and it is also likely associated with tobacco use, thus meeting the criteria for a confounder.

Mediator:

Table 3: Model 3: Logistic Regression Model with Tobacco, Family History, Age, and LDL Predicting Coronary Heart Disease

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.19831	0.75022	-5.596	2.19e-08 ***
tobacco	0.08767	0.02587	3.389	0.000700 ***
famhist_f1	0.94778	0.22444	4.223	2.41e-05 ***
age_f2	1.57619	0.78524	2.007	0.044720 *
age_f3	1.95055	0.77302	2.523	0.011626 *
age_f4	1.84772	0.77398	2.387	0.016973 *
age_f5	2.55969	0.76372	3.352	0.000803 ***
ldl	0.16724	0.05439	3.075	0.002108 **

Note: Null deviance: 596.11 on 461 degrees of freedom. Residual deviance: 484.36 on 454 degrees of freedom. AIC: 500.36. Number of Fisher Scoring iterations: 6.

Table 4: Model 4: Logistic Regression Model with Tobacco, Age, LDL, and Interaction between LDL and Family History Predicting Coronary Heart Disease

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.51715	0.77012	-4.567	4.95e-06 ***
tobacco	0.09169	0.02628	3.489	0.000485 ***
age_f2	1.61631	0.78582	2.057	0.039702 *
age_f3	2.03148	0.77429	2.624	0.008699 **
age_f4	1.89128	0.77583	2.438	0.014778 *
age_f5	2.65150	0.76542	3.464	0.000532 ***
ldl	0.01104	0.07432	0.149	0.881874
famhist_f1	-0.84505	0.62627	-1.349	0.177232
ldl:famhist_f1	0.35605	0.11712	3.040	0.002366 **

Note: Null deviance: 596.11 on 461 degrees of freedom. Residual deviance: 474.35 on 453 degrees of freedom. AIC: 492.35. Number of Fisher Scoring iterations: 6.

Those with a family history of Coronary heart disease(CHD) could have a stronger chance of having CHD due to different ldl levels. Thus, we introduce interaction term, ldl* famhist_f.

ANOVA Likelihood Ratio Test for Model Comparison

Model 1: y ~ sbp + tobacco + ldl + adiposity + famhist_f + typea + obesity + alcohol+age_f
Model 2: y ~ tobacco + ldl + famhist_f + typea + obesity + age_f
Model 3: y ~ tobacco + ldl + famhist_f + typea + age_f
Model 4: y~tobacco + ldl + famhist_f + typea + age_f + ldl*famhist_f

Table 5: Analysis of Deviance Table for Model Comparison

Model	Resid. Df	Resid. Dev	Df	Deviance
Model 1	452	474.24		
Model 2	449	470.94	3	3.3003

Note: $\Pr(>\chi) = 0.3476$.

Table 6: Analysis of Deviance Table for Model Comparison

Model	Resid. Df	Resid. Dev	Df	Deviance
Model 1	452	474.24		
Model 2	453	475.96	-1	-1.7196

Note: $\Pr(>\chi) = 0.1897$.

Table 7: Analysis of Deviance Table for Model Comparison

Model	Resid. Df	Resid. Dev	Df	Deviance
Model 1	453	475.96		
Model 2	452	466.29	1	9.6773

Note: $\Pr(>\chi) = 0.001866$. **Significant at 0.01 level.

p-value > 0.05 so implies both models are not significantly different. So not including the 3 predictors is the same as including them, Therefore, we can move forward with the simpler model_2.

Also obesity predictor has a p-value<0.05. We can remove it and check.

So we can indeed remove the Obesity predictor since p-value > 0.05.

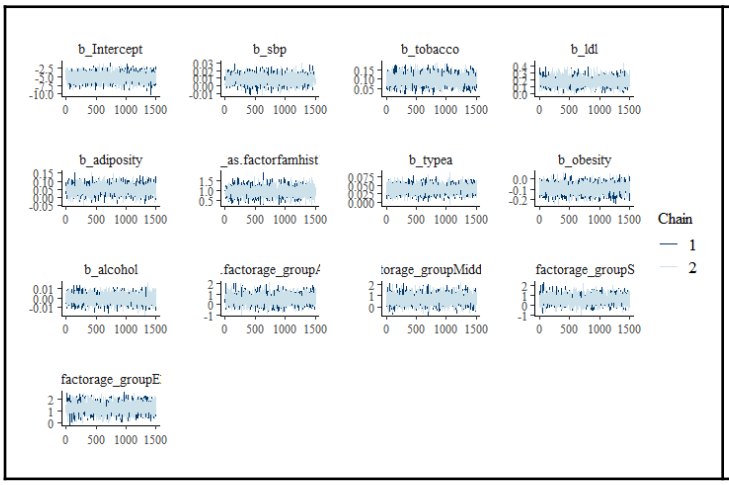
We reject the null hypothesis that model_3 fits the same as model_4.
So we must include the interaction term.

10-fold Cross-validation Result:

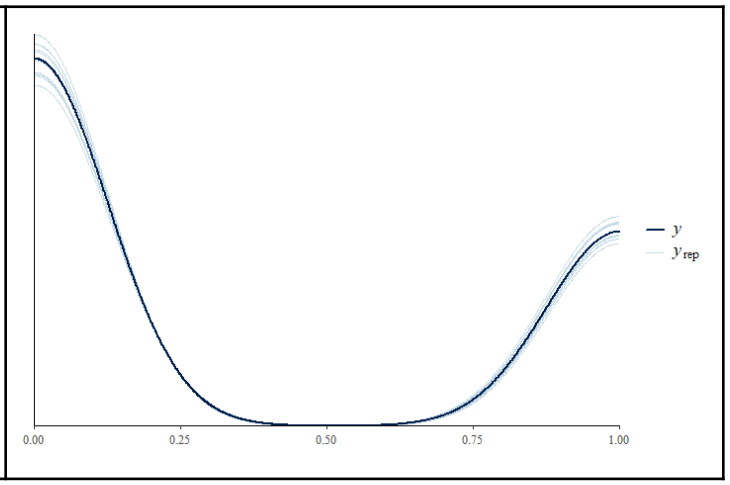
Table 8: 10-Fold Cross-Validation Results

Model	Accuracy	Kappa
$y \sim tobacco + ldl + famhist\_f + typea + age\_f + ldl : famhist\_f$	0.7314986	0.3719403
$y \sim .$	0.7141073	0.3384958

Stan Plot of predictors



Posterior Predictive Check



The Stan plot for all chains shows a good parameter mix and appears as a ‘hairy caterpillar’ which suggests good mixing and convergence.

"y" represents the observed data, while "y_rep" represents replicated data generated from the model's predictive distribution. The fact that the lines are close to each other suggests that the replicated data closely matches the observed data, indicating a good model fit.

Updating different Prior values to improve out-of-sample predictive power

Original model:

```
  set_prior("normal(0, 2.5)", class = "Intercept"),   set_prior("normal(0, 1)", class = "b") )
```

A model with more conservative priors (i.e., more variance)

```
fit_conservative <- update(
  fit_original,
  prior = c(
    set_prior("normal(0, 5)", class = "Intercept"), set_prior("normal(0, 2.5)", class = "b")))
```

A model with more informative priors

```
fit_informative <- update(
  fit_original,
  prior = c(  set_prior("normal(0, 1)", class = "Intercept"), set_prior("normal(0, 0.5)", class = "b")))
```

```
                elpd_diff se_diff
fit_conservative  0.0       0.0
fit_original     -1.2       1.2
```

```
                elpd_diff se_diff
fit_original     0.0       0.0
fit_informative -1.6       1.0
```

A higher ELPD indicates a model with better out-of-sample predictive accuracy. So, the fit_conservative model performs better than fit_original which has an elpd_diff of -1.2. This means that fit_original has a lower predictive accuracy compared to fit_conservative on log-scale. And when we compare fit_original model with fit_informative, the fit_original performs better which implies it has better out-of-sample predictive power.