# Transporting the framingham risk prediction model for use in a new target population

Kerry Ye

2023-11-30

**Abstract**

Objective: The source data often cannot be treated as a random sample from the target population due to differences in covariate distributions between the two populations. In this project, our goal is to apply both the conventional and new extension of the brier score methods to: 1) Evaluate the performance of the risk score model in the source population (Framingham Heart Study data) that was developed using the same data; 2) Evaluate the performance of the risk score model in the National Health and Nutrition Examination Survey (NHANES) study, which was developed using Framingham data; 3) Evaluate the performance of the risk score model using simulated NHANES data developed from the Framingham data.

Methods: We employ a logistic regression model, developed from the Framingham dataset, which is trained using the training set. This model is used to predict the risk of cardiovascular disease, with each version of the model stratified by sex. Additionally, we utilize logistic regression to estimate the inverse-odds weighting estimator by calculating $\Pr[S = 1 \mid X, D_{\text{test},i} = 1]$, and assessed the performance of each sex-stratified risk score prediction model using modified Brier Score when the target population is different from the source populatioon.

Results: The Brier scores for the women's risk score prediction model were consistently lower than those for the men's model. Intriguingly, the model performance, when developed and evaluated on the same dataset (Framingham), showed higher Brier scores compared to when the model was applied to new target populations (actual and simulated NHANES). Moreover, the Brier score for the men's prediction model in the composite dataset with Framingham and Simulated NHANES was higher than that in the composite dataset with Framingham and actual NHANES. Conversely, the performance of the women's prediction model was similar in these two datasets.

## Introduction

In collaboration with Dr. Jon Steingrimsson, this project aims to investigate the performance of a prediction model in a target population that differs from the population originally used for model development. Typically, the data used for building prediction models, known as source study data, are derived from randomized trials, large observational databases. However, this data often cannot be treated as a random sample from the target population due to differences in covariate distributions between the two populations.

Consider a setup where both covariate and outcome data are available in the source population, but only covariate data is available in the target population. Consequently, using traditional performance measures to evaluate a prediction model developed in the source population and applied to the target population can be problematic, as outcome data is not available from the target population. This limitation makes the prediction model less applicable in the target population, and the model's performance measured in the source population may not reflect its performance in the target population.

Recently, several methods have been developed to evaluate the performance of transporting a prediction model for use in a new target population. In this project, our goal is to apply these methods to: 1) Evaluate the performance of the risk score model in the source population (Framingham Heart Study data) that was developed using the same data; 2) Evaluate the performance of the risk score model in the National Health and Nutrition Examination Survey (NHANES) study, which was developed using Framingham data; 3) Evaluate

the performance of the risk score model using simulated NHANES data developed from the Framingham data. As the framingham includes data NHANES does not include outcome information, this involves using composite data from both NHANES and the Framingham Heart Study to examine the prediction model's transportation performance.

# Methodology

## Data Source

Framingham: The Framingham Heart Study is a long-term investigation of cardiovascular disease among a population in Framingham, Massachusetts. This study was the first of its kind to prospectively examine cardiovascular disease, identifying the concept of risk factors and their combined effects. Initiated in 1948, the study enrolled 5,209 subjects. Participants have been examined biennially, and all subjects are continuously monitored for cardiovascular outcomes. Clinical examination data includes cardiovascular disease risk factors and markers, such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, echocardiography, and medication use. The enclosed dataset represents a subset of the data collected in the Framingham study, comprising laboratory, clinical, questionnaire, and adjudicated event data on 4,434 participants.

NHANES: The National Health and Nutrition Examination Survey (NHANES) is a national survey conducted to monitor the health and nutritional status of both adults and children in the United States. What sets NHANES apart is its combination of interviews and physical examinations. The survey is administered by a part of the Centers for Disease Control and Prevention (CDC). The first NHANES program was initiated in 1960. Since 1999, the survey has examined about 5,000 individuals yearly in 15 different counties across the country. The data collected by NHANES is instrumental in providing vital health statistics.

## Model

The prediction model presented below is a logistic regression model designed to assess the risk factors for cardiovascular disease (CVD) using data from the Framingham Heart Study. This model is fitted separately for men and women for subsequent analysis:

$$\begin{aligned}\text{logit}(\Pr(\text{CVD}=1)) = \beta_0 &+ \beta_1 \log(\text{HDLC}) + \beta_2 \log(\text{TOTCHOL}) \\ &+ \beta_3 \log(\text{AGE}) + \beta_4 \log(\text{SYSBP\_UT}+1) \\ &+ \beta_5 \log(\text{SYSBP\_T}+1) + \beta_6 \text{CURSMOKE} \\ &+ \beta_7 \text{DIABETES}\end{aligned}$$

The outcome variable of interest is a binary indicator indicating whether cardiovascular disease occurred (CVD=1) or did not occur (CVD=0) during the follow-up period, and the logit function is used to denote the logistic transformation, which is the natural log of the odds of the event (CVD = 1 in this case).

The model includes several continuous variables: High-Density Lipoprote in Cholesterol (HDLC) in mg/dL, Serum Total Cholesterol (TOTCHOL) measured in mg/dL, and age at exam (AGE) in years. Additionally, the model accounts for the use of anti-hypertensive medication (BPMEDS), distinguishing between those who currently use these medications and those who do not. Systolic Blood Pressure (SYSBP) taken as the mean of the last two of three measurements in mmHg. SYSBP is further bifurcated into two variables: SYSBP_UT and SYSBP_T. SYSBP_UT represents the systolic blood pressure for individuals not currently using anti-hypertensive medication, while SYSBP_T represents the systolic blood pressure for individuals who are on such medication. These variables undergo logarithmic transformation to approximate normality, addressing skewness in the data.

There are also two categorical variables: Diabetes (DIABETES), where 0 indicates non-diabetic and 1 indicates diabetic based on the criteria of the first exam treated or a casual glucose level of 200 mg/dL or more; and

current cigarette smoking status (CURSMOKE), with 0 representing non-smokers and 1 representing current smokers.

### Metrics

As traditional performance metrics are not suitable for evaluating the use of prediction models developed from source data in a target population, Steingrimsson, et al. proposed a modified version of the Brier Score[1].

Let S be an indicator for the population from which data are obtained, with S = 1 for the source population and S = 0 for the target population. Let $D_{test}$ denote the test set indicator, where $D_{\text{test}} = 1$ if an observation is in either the source or target test set. Denote n = $n_{\text{source}}$ + $n_{\text{target}}$ as the sample size of the composite dataset consisting of data from both the source and target population samples. Use the function $g_{\hat{\beta}}(X)$ to denote the "fitted" model with estimated parameter $\hat{\beta}$.

With the inverse-odds weighting estimator, the Brier Score for the target population is defined as follows:

$$\hat{\psi}_{\hat{\beta}} = \frac{\sum_{i=1}^{n} I(S_i = 1, D_{\text{test},i} = 1)\hat{o}(X_i)(Y_i - g_{\hat{\beta}}(X)))^2}{\sum_{i=1}^{n} I(S_i = 0, D_{\text{test},i} = 1)}$$

,

where the function $\hat{o}(X_i$ is the inverse-odds weighting estimator in the test set, defined as $\frac{\Pr[S=0|X,D_{\text{test},i}=1]}{\Pr[S=1|X,D_{\text{test},i}=1]}$.

## Analysis

We divided the project goal into three analyses. Since we only consider complete cases in this project, we omitted the missing values present in the NHANES data.

### Analysis I

The first goal of this project is to evaluate the performance of the risk score model in the source population, using data from the Framingham Heart Study that was also utilized to develop the model.

Initially, we divided the Framingham data into two subsets: one for men and another for women. For each dataset, we partitioned the data into training and testing sets by randomly selecting 70% of the data for training and the remaining 30% for testing. We then used the training data to fit the prediction model as previously described and applied the testing data to evaluate its performance. Since both the source and target populations are derived from the Framingham study, we employed traditional evaluation metrics, specifically the Brier Score, to assess the performance of the risk score prediction model. The conventional Brier Score represents the average squared difference between the predicted probabilities and the actual outcomes. The mathematical formula for the Brier score is: Brier score = $\frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$, where $f_i$ is the predicted probability for the i-th event, $o_i$ represents the observed outcome of the i-th event (0, 1), and $N$ is the total number of prediction. In this analysis, we calculate two Brier scores to evaluate the risk score prediction model, stratified by sex.

### Analysis II

The second goal of this project is to evaluate the performance of the risk score model in the NHANES study, which was developed using data from the Framingham Heart Study.

Initially, we created a composite dataset by combining the Framingham and NHANES datasets. It's important to note that the Framingham data includes both outcome and covariate information, whereas the NHANES dataset contains only covariate data. We then stratified the composite dataset into two groups: men and women. For each group, we partitioned the data into training and testing sets by randomly selecting 70% of the data for training and the remaining 30% for testing. The training data was used to fit the prediction model as previously described, and the testing data was used for performance evaluation. Since the NHANES data

is not a random sample from the Framingham data, and these two populations can exhibit very different data distributions, we employed the modified version of the Mean Squared Error (MSE) as previously described. This method evaluates the performance of the risk score prediction model developed from the Framingham data when applied to the NHANES study. The inverse odds weighting estimator is estimated using logistic regression to determine $\Pr[S = 1 \mid X, D_{\text{test},i} = 1]$ in the test set. Eventually, we calculate two Brier scores to evaluate the risk score prediction model, stratified by sex.

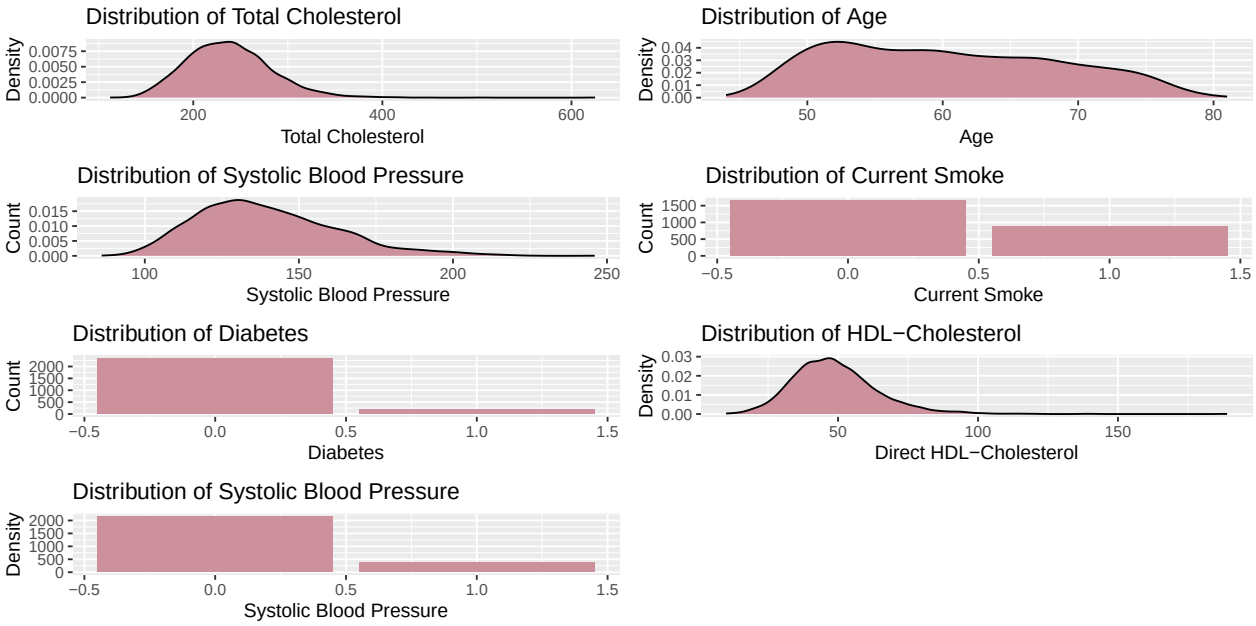
## Analysis III

We would discuss the simulation process by using the following ADEMP framework:

**Aims:** The third goal of this project is to evaluate the performance of the risk score model using simulated NHANES data, where the model is developed from the Framingham data.

**Data-Generating Mechanisms:** The data-generating process for the simulated NHANES data is informed by both the distribution shape from the Framingham dataset and the summary statistics table from the actual NHANES dataset.

Below are the distribution plots for each covariate in the Framingham dataset used for risk score prediction model development:




As we can observe, continuous variables such as TOTCHOL, SYSBP, and HDLC are normally distributed, with some skewness to the left. However, the variable AGE, which always has high density and can only be a positive number, should use a uniform distribution instead of a normal distribution for NHANES data generation. Other categorical variables like DIABETES, CURSMOKE, and BPMEDS can be generated using a binomial distribution.

We have recreated the summary statistics table from the NHANES data, focusing on complete case analysis for this project. This table contains mean and standard deviation information stratified by sex for each covariate in the actual NHANES data.

In the given framework, we consider two data-generating mechanisms, setting the seed to 1 and iterating each simulation 500 times.

For men, the NHANES data are simulated for $n_{\text{target}} = 2105$, representing the male survey participants in the National Health and Nutritional Examination Study. The sex indicator is set as 1, indicating male participants. The continuous variable HDLC for each participant is generated from a normal distribution as

Table 1: Summary Statistics Stratified by SEX

| | 1 | 2 | p |
|---|---|---|---|
| n | 2105 | 2205 | |
| SYSBP (mean (SD)) | 126.44 (16.83) | 123.70 (20.36) | <0.001 |
| SEX (mean (SD)) | 1.00 (0.00) | 2.00 (0.00) | <0.001 |
| AGE (mean (SD)) | 50.15 (18.83) | 48.90 (18.57) | 0.029 |
| BMI (mean (SD)) | 29.19 (6.25) | 29.84 (7.96) | 0.003 |
| HDLC (mean (SD)) | 48.11 (13.59) | 58.10 (15.68) | <0.001 |
| CURSMOKE (mean (SD)) | 0.20 (0.40) | 0.14 (0.35) | <0.001 |
| BPMEDS (mean (SD)) | 0.30 (0.46) | 0.29 (0.45) | 0.584 |
| TOTCHOL (mean (SD)) | 183.10 (41.65) | 190.51 (41.20) | <0.001 |
| DIABETES (mean (SD)) | 0.18 (0.38) | 0.12 (0.33) | <0.001 |
| SYSBP_UT (mean (SD)) | 86.46 (57.73) | 83.64 (55.42) | 0.101 |
| SYSBP_T (mean (SD)) | 39.98 (62.19) | 40.07 (63.62) | 0.964 |

$\mathcal{N}(48.11, sd = 13.59)$; TOTCHOL, is generated as $\mathcal{N}(183.10, sd = 41.65)$; AGE is derived from a uniform distribution as $Unif(18, 80)$; SYSBP, is generated as $\mathcal{N}(126.44, sd = 16.83)$, with the binary indicator BPMEDS generated from a binomial distribution as $Binom(1, 0.30)$. The variables SYSBP_UT and SYSBP_T represent the systolic blood pressure for individuals not on or on anti-hypertensive medication, respectively. For all these continuous variables, if any generated value is below 0, it is changed to 1. The other two categorical variables, DIABETES and CURSMOKE, are generated as $Binom(1, 0.18)$ and $Binom(1, 0.20)$, respectively.

For women, the NHANES data are simulated on $n_{\text{target}} = 2205$, representing the female survey participants. The sex indicator is set as 2 for female participants. HDLC for each participant is generated as $\mathcal{N}(58.10, sd = 15.68)$; TOTCHOL, is generated as $\mathcal{N}(190.51, sd = 41.20)$; AGE, is derived from $Unif(18, 80)$; SYSBP, is derived from $\mathcal{N}(123.70, sd = 20.36)$, with BPMEDS generated as $Binom(1, 0.29)$. SYSBP_UT and SYSBP_T are defined similarly as for men. For all continuous variables, values below 0 are adjusted to 1. The other categorical variables, DIABETES and CURSMOKE, are generated generated from $Binom(1, 0.12)$ and $Binom(1, 0.14)$, respectively.

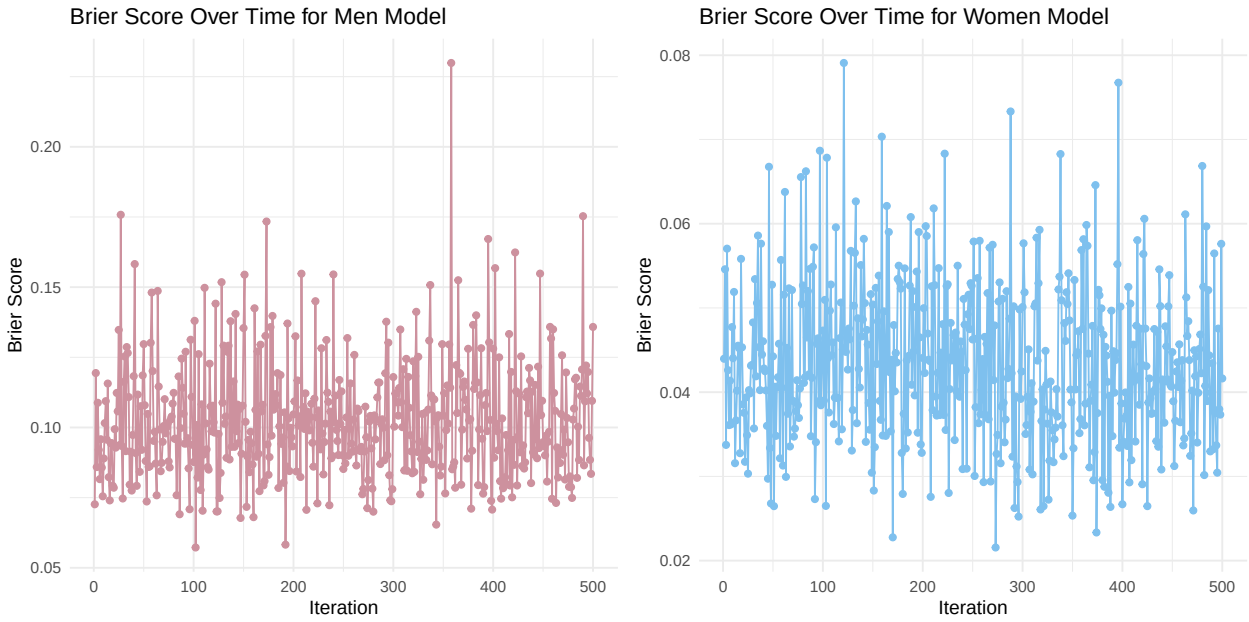
**Estimands:** Our estimand is the Brier Score defined in the metrics sub-section that specifically designed for the evaluate the performance of the prediction developed in the framingham Study in the simulated NHANES dataset.

**Methods:** The method employed in this simulation is to combine the Framingham data with the simulated NHANES data to create a composite dataset. We then employ a logistic regression model, developed from the Framingham dataset, which is trained using the training set. This model is used to predict the risk of cardiovascular disease, with each version of the model stratified by sex. Additionally, we utilize logistic regression to estimate the inverse-odds weighting estimator by calculating $\Pr[S = 1 \mid X, D_{\text{test},i} = 1]$.

**Performance Measures:** We assessed the performance of each sex-stratified risk score prediction model in the target population by calculating the average of the modified Brier Score using the test set across 500 simulations.

# Results

In analysis III, the brier score simulated in 500 times for each risk score prediction model stratfied by sex is shown below:




In Analysis III, the Brier scores for each sex-stratified risk score prediction model were simulated 500 times. We observed that the simulations generally produced stable Brier scores within a relatively small range for both the men's and women's prediction models. However, a few iterations resulted in extremely small or large Brier score values. Overall, the Brier scores estimated for the women's risk score prediction model were consistently smaller than those estimated for the men's model.

In our study, we consolidated the Brier score results from Analyses I, II, and III into a comprehensive performance results table. The Brier score, a metric used to evaluate the accuracy of probabilistic predictions, ranges from 0 for perfect predictions to 1 for the least accurate.

Table 2: Brier Scores Stratified by Sex

| | Framingham | Framingham + NHANES | Framingham + Simulated NHANES |
|---|---|---|---|
| Men Model | 0.1916928 | 0.0771307 | 0.1024022 |
| Women Model | 0.1175866 | 0.0443940 | 0.0433863 |

In Analysis I, we employed the regular Brier score to evaluate the performance of the risk score prediction model. This was appropriate since both the source and target populations were derived from the Framingham study. However, in Analyses II and III, we used a modified version of the Brier score. This adjustment was necessary because both the actual NHANES and the simulated NHANES data are not random samples from the Framingham data, and these populations can exhibit significantly different data distributions.

Across all analyses, it was observed that the Brier scores for the women's risk score prediction model were consistently lower than those for the men's model. Intriguingly, the model performance, when developed and evaluated on the same dataset (Framingham), showed higher Brier scores compared to when the model was applied to new target populations (actual and simulated NHANES). This suggests that the composite data (Framingham + NHANES and Framingham + Simulated NHANES) provide more accurate predictions of patients' cardiovascular disease risk than the Framingham data alone.

Moreover, the Brier score for the men's prediction model in the composite dataset with Framingham and Simulated NHANES was higher than that in the composite dataset with Framingham and actual NHANES. This difference indicates a potential bias in the simulated data, which was based solely on summary statistics, compared to the actual NHANES data. Conversely, the performance of the women's prediction model was similar in these two datasets, suggesting a consistent predictive accuracy across these two data compositions.

## Discussion

Overall, all three analyses demonstrate that the various data compositions—Framingham, Framingham combined with actual NHANES, and Framingham combined with simulated NHANES—exhibit high accuracy and strong performance in models used for predicting cardiovascular disease risk. This holds true whether the models are applied within the same dataset or transported to a new target population. Notably, the model performance of risk score predictions developed and evaluated using the same dataset (Framingham) was found to be inferior compared to Analyses II and III. This discrepancy could be attributed to the smaller sample size of the source data and its potentially more homogeneous nature compared to the composite datasets, which blend both source and target populations. The findings from this project underscore the critical importance of employing appropriate evaluation metrics when transporting models developed from a source population to assess their performance in a target population. The results also highlight the limitations inherent in using simulated data. Additionally, the project reveals that predictive models may perform differently across sexes, suggesting the need for sex-specific considerations in model development and evaluation.

This project has several limitations. First, it relies solely on one evaluation metric, which may not accurately reflect the true performance of the risk score model. Researchers should consider using the modified version of the Area Under the Curve (AUC), as proposed by Li, et al.[3], for a more comprehensive evaluation. Second, the project assumes that covariates are independent of each other and does not account for potential correlations between different covariates in the prediction model during the data generation process for the simulation. Researchers may want to consider a multivariate normal distribution if suspected correlations are found in the dataset.

# References

[1] Steingrimsson, Jon A., et al. "Transporting a prediction model for use in a new target population." American Journal of Epidemiology 192.2 (2023): 296-304.

[2] Morris, Tim P., Ian R. White, and Michael J. Crowther. "Using simulation studies to evaluate statistical methods." Statistics in medicine 38.11 (2019): 2074-2102.

[3] Li, Bing, et al. "Estimating the area under the ROC curve when transporting a prediction model to a target population." Biometrics 79.3 (2023): 2382-2393.

## Code Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(message = F)
knitr::opts_chunk$set(warning = F)
library(riskCommunicator)
library(tidyverse)
library(tableone)

data("framingham")

# The Framingham data has been used to create models for cardiovascular risk.
# The variable selection and model below are designed to mimic the models used
# in the paper General Cardiovascular Risk Profile for Use in Primary Care
# This paper is available (cvd_risk_profile.pdf) on Canvas.

framingham_df <- framingham %>% select(c(CVD, TIMECVD, SEX, TOTCHOL, AGE,
                                         SYSBP, DIABP, CURSMOKE, DIABETES, BPMEDS,
                                         HDLC, BMI))
framingham_df <- na.omit(framingham_df)

CreateTableOne(data=framingham_df, strata = c("SEX"))

# Get blood pressure based on whether or not on BPMEDS
framingham_df$SYSBP_UT <- ifelse(framingham_df$BPMEDS == 0,
                                 framingham_df$SYSBP, 0)
framingham_df$SYSBP_T <- ifelse(framingham_df$BPMEDS == 1,
                                framingham_df$SYSBP, 0)

# Looking at risk within 15 years - remove censored data
dim(framingham_df)
framingham_df <- framingham_df %>%
  filter(!(CVD == 0 & TIMECVD <= 365*15)) %>%
  select(-c(TIMECVD))
dim(framingham_df)

# Filter to each sex
framingham_df_men <- framingham_df %>% filter(SEX == 1)
framingham_df_women <- framingham_df %>% filter(SEX == 2)

# Fit models with log transforms for all continuous variables
mod_men <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                 log(SYSBP_T+1)+CURSMOKE+DIABETES,
      data= framingham_df_men, family= "binomial")


mod_women <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                   log(SYSBP_T+1)+CURSMOKE+DIABETES,
               data= framingham_df_women, family= "binomial")


# The NHANES data here finds the same covariates among this national survey data
library(nhanesA)
```

```r
# blood pressure, demographic, bmi, smoking, and hypertension info
bpx_2017 <- nhanes("BPX_J") %>%
  select(SEQN, BPXSY1 ) %>%
  rename(SYSBP = BPXSY1)
demo_2017 <- nhanes("DEMO_J") %>%
  select(SEQN, RIAGENDR, RIDAGEYR) %>%
  rename(SEX = RIAGENDR, AGE = RIDAGEYR)
bmx_2017 <- nhanes("BMX_J") %>%
  select(SEQN, BMXBMI) %>%
  rename(BMI = BMXBMI)
smq_2017 <- nhanes("SMQ_J") %>%
  mutate(CURSMOKE = case_when(SMQ040 %in% c(1,2) ~ 1,
                              SMQ040 == 3 ~ 0,
                              SMQ020 == 2 ~ 0)) %>%
  select(SEQN, CURSMOKE)
bpq_2017 <- nhanes("BPQ_J") %>%
  mutate(BPMEDS = case_when(
    BPQ020 == 2 ~ 0,
    BPQ040A == 2 ~ 0,
    BPQ050A == 1 ~ 1,
    TRUE ~ NA )) %>%
  select(SEQN, BPMEDS)
tchol_2017 <- nhanes("TCHOL_J") %>%
  select(SEQN, LBXTC) %>%
  rename(TOTCHOL = LBXTC)
hdl_2017 <- nhanes("HDL_J") %>%
  select(SEQN, LBDHDD) %>%
  rename(HDLC = LBDHDD)
diq_2017 <- nhanes("DIQ_J") %>%
  mutate(DIABETES = case_when(DIQ010 == 1 ~ 1,
                              DIQ010 %in% c(2,3) ~ 0,
                              TRUE ~ NA)) %>%
  select(SEQN, DIABETES)

# Join data from different tables
df_2017 <- bpx_2017 %>%
  full_join(demo_2017, by = "SEQN") %>%
  full_join(bmx_2017, by = "SEQN") %>%
  full_join(hdl_2017, by = "SEQN") %>%
  full_join(smq_2017, by = "SEQN") %>%
  full_join(bpq_2017, by = "SEQN") %>%
  full_join(tchol_2017, by = "SEQN") %>%
  full_join(diq_2017, by = "SEQN")

CreateTableOne(data = df_2017, strata = c("SEX"))


# Filter to each sex
framingham_df_men <- framingham_df %>% filter(SEX == 1)
framingham_df_women <- framingham_df %>% filter(SEX == 2)
#make this example reproducible
set.seed(1)
```

```r
#use 70% of dataset as training set and 30% as test set
sample_men <- sample(c(TRUE, FALSE), nrow(framingham_df_men), replace=TRUE, prob=c(0.7,0.3))
train_men  <- framingham_df_men[sample_men, ]
test_men   <- framingham_df_men[!sample_men, ]

sample_women <- sample(c(TRUE, FALSE), nrow(framingham_df_women), replace=TRUE, prob=c(0.7,0.3))
train_women  <- framingham_df_women[sample_women, ]
test_women   <- framingham_df_women[!sample_women, ]
mod_train_men <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                 log(SYSBP_T+1)+CURSMOKE+DIABETES,
      data= train_men, family= "binomial")

mod_train_women <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                   log(SYSBP_T+1)+CURSMOKE+DIABETES,
               data= train_women, family= "binomial")
x_test_men <- subset(test_men, select=-c(CVD))
y_test_men <- test_men$CVD
x_test_women <- subset(test_women, select = -c(CVD))
y_test_women <- test_women$CVD
predicted_men <- predict(mod_train_men, test_men, type = "response")
predicted_women <- predict(mod_train_women, test_women, type = "response")
library(pROC)
roc_men <- roc(predictor = predicted_men, response = y_test_men)
roc_coord <- coords(roc = roc_men, x = "best")
auc(roc_men)

roc_women <- roc(predictor = predicted_women, response = y_test_women)
roc_coord <- coords(roc = roc_women, x = "best")
auc(roc_women)
library(DescTools)
framingham_bs_men <- BrierScore(y_test_men, predicted_men)
framingham_bs_women <-BrierScore(y_test_women, predicted_women)
# Get blood pressure based on whether or not on BPMEDS
df_2017$SYSBP_UT <- ifelse(df_2017$BPMEDS == 0, df_2017$SYSBP, 0)

df_2017$SYSBP_T <- ifelse(df_2017$BPMEDS == 1, df_2017$SYSBP, 0)
framingham_df_com <- subset(framingham_df, select = -c(DIABP))
framingham_df_com$S <- 1

df_2017_com <- subset(df_2017, select = -c(SEQN))
df_2017_com$S <- 0
df_2017_com <- na.omit(df_2017_com)
df_2017_com$CVD <- NA
comp_dt <- rbind(framingham_df_com, df_2017_com)
comp_dt$CURSMOKE <- as.factor(comp_dt$CURSMOKE)
comp_dt$DIABETES <- as.factor(comp_dt$DIABETES)
comp_dt_men <- comp_dt %>% filter(SEX == 1)
comp_dt_women <- comp_dt %>% filter(SEX == 2)

#make this example reproducible
set.seed(1)

#use 70% of dataset as training set and 30% as test set
```

```r
comp_sample_men <- sample(c(TRUE, FALSE), nrow(comp_dt_men), replace=TRUE, prob=c(0.7,0.3))
train_comp_men  <- comp_dt_men[comp_sample_men, ]
test_comp_men   <- comp_dt_men[!comp_sample_men, ]


comp_sample_women <- sample(c(TRUE, FALSE), nrow(comp_dt_women), replace=TRUE, prob=c(0.7,0.3))
train_comp_women  <- comp_dt_women[comp_sample_women, ]
test_comp_women   <- comp_dt_women[!comp_sample_women, ]
mod_train_comp_men <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES,
      data= train_comp_men, family= "binomial")

mod_train_comp_women <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                  log(SYSBP_T+1)+CURSMOKE+DIABETES,
               data= train_comp_women, family= "binomial")
x_test_comp_men <- subset(test_comp_men, select=-c(CVD))
y_test_comp_men <- test_comp_men$CVD
x_test_comp_women <- subset(test_comp_women, select = -c(CVD))
y_test_comp_women <- test_comp_women$CVD
test_comp_men$predicted_comp_men <- predict(mod_train_comp_men, x_test_comp_men, type = "response")
test_comp_women$predicted_comp_women <- predict(mod_train_comp_women, x_test_comp_women , type = "respo
fit_men <- glm(S ~ log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES, data = test_comp_men, family = "binomial")
pred.prob.men <- predict(fit_men, type = "response")
test_comp_men$inv_odd_weight_men <- (1-pred.prob.men)/pred.prob.men

fit_women <- glm(S ~ log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES, data = test_comp_women, family = "binomial")
pred.prob.women <- predict(fit_women, type = "response")
test_comp_women$inv_odd_weight_women <- (1-pred.prob.women)/pred.prob.women

#s_d_indicator_men <- ifelse(comp_dt_men$S == 1 & comp_sample_men == 0, 1, 0)
brierscore_men <- sum(test_comp_men$inv_odd_weight_men[test_comp_men$S==1] * (test_comp_men$CVD[test_co

brierscore_women <- sum(test_comp_women$inv_odd_weight_women[test_comp_women$S==1] * (test_comp_women$C
library(gridExtra)
a1 <- framingham_df_com %>%
ggplot(aes(x = TOTCHOL)) + geom_density(fill = "pink3") +
  labs(title = "Distribution of Total Cholesterol", x = "Total Cholesterol", y = "Density")


a2 <- framingham_df_com %>%
  ggplot(aes(x = AGE)) + geom_density(fill = "pink3") +
  labs(title = "Distribution of Age", x = "Age", y = "Density")

a3 <- framingham_df_com %>%
  ggplot(aes(x = SYSBP)) + geom_density(fill = "pink3") +
  labs(title = "Distribution of Systolic Blood Pressure", x = "Systolic Blood Pressure", y = "Count")



a4 <- framingham_df_com %>%
  ggplot(aes(x = CURSMOKE)) + geom_bar(fill = "pink3") +
```

```
  labs(title = "Distribution of Current Smoke", x = "Current Smoke", y = "Count")

a5 <- framingham_df_com %>%
  ggplot(aes(x = DIABETES)) + geom_bar(fill = "pink3")+
  labs(title = "Distribution of Diabetes", x = "Diabetes", y = "Count")


a6 <- framingham_df_com  %>%
ggplot(aes(x = HDLC)) + geom_density(fill = "pink3") +
  labs(title = "Distribution of HDL-Cholesterol", x = "Direct HDL-Cholesterol", y = "Density")

a7 <- framingham_df_com  %>%
  ggplot(aes(x = BPMEDS)) + geom_bar(fill = "pink3") +
  labs(title = "Distribution of Systolic Blood Pressure", x = "Systolic Blood Pressure", y = "Density")

grid.arrange(a1, a2, a3, a4, a5, a6, a7, ncol=2)
tableone <- CreateTableOne(data = df_2017_com[,-12], strata = c("SEX"))
tableone
framingham_df_women <- subset(framingham_df_women, select=-c(DIABP))
framingham_df_women$S <- 1
framingham_df_men <- subset(framingham_df_men, select=-c(DIABP))
framingham_df_men$S <- 1
sim_brierscore_men <- c()
set.seed(1)

for (i in 1:500){

n_men <- 2105
SYSBP_men <- rnorm(n_men, 126.44, 16.83)
SEX_men <- rep(1, n_men)
AGE_men <- runif(n_men, 18, 80)
BMI_men <- rnorm(n_men, 26.16, 7.63)
BMI_men <- ifelse(BMI_men < 0, 1, BMI_men)
HDLC_men <- rnorm(n_men, 48.11, 13.59)
HDLC_men <- ifelse(HDLC_men < 0, 1, HDLC_men)
CURSMOKE_men <- rbinom(n_men, 1, 0.20)
BPMEDS_men <- rbinom(n_men, 1, 0.30)
SYSBP_UT_men <- ifelse(BPMEDS_men == 0, SYSBP_men, 0)
SYSBP_T_men <- ifelse(BPMEDS_men == 1, SYSBP_men, 0)
TOTCHOL_men <- rnorm(n_men, 183.10, 41.65)
TOTCHOL_men <- ifelse(TOTCHOL_men < 0, 1, TOTCHOL_men)
DIABETES_men <- rbinom(n_men, 1, 0.18)
CVD_men <- NA
S_men <- 0

sim_df_2017_men <- data.frame(CVD_men, SYSBP_men, SEX_men, AGE_men, BMI_men, HDLC_men, CURSMOKE_men, BPI
names(sim_df_2017_men) <- c("CVD", "SYSBP", "SEX", "AGE", "BMI", "HDLC", "CURSMOKE", "BPMEDS", "SYSBP_U
sim_comp_men <- rbind(framingham_df_men, sim_df_2017_men)

#use 70% of dataset as training set and 30% as test set
sim_comp_sample_men <- sample(c(TRUE, FALSE), nrow(sim_comp_men), replace=TRUE, prob=c(0.7,0.3))
sim_train_comp_men  <- sim_comp_men[sim_comp_sample_men, ]
sim_test_comp_men   <- sim_comp_men[!sim_comp_sample_men, ]
```

```
mod_sim_train_comp_men <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                 log(SYSBP_T+1)+CURSMOKE+DIABETES,
      data= sim_train_comp_men, family= "binomial")

x_test_sim_comp_men <- subset(sim_test_comp_men, select=-c(CVD))
y_test_sim_comp_men <- sim_test_comp_men$CVD

sim_test_comp_men$predicted_comp_men <- predict(mod_sim_train_comp_men, sim_test_comp_men, type = "respo

sim_fit_men <- glm(S ~ log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                 log(SYSBP_T+1)+CURSMOKE+DIABETES, data = sim_test_comp_men, family = "binomial")
sim.pred.prob.men <- predict(sim_fit_men, type = "response")
sim_test_comp_men$inv_odd_weight_men <- (1-sim.pred.prob.men)/sim.pred.prob.men

sim_brierscore_men[i] <- sum(sim_test_comp_men$inv_odd_weight_men[sim_test_comp_men$S==1] * (sim_test_co

}

hist(df_2017_com$SYSBP[df_2017_com$SEX==1])
hist(SYSBP_men)
hist(df_2017_com$AGE[df_2017_com$SEX==1])
hist(AGE_men)
hist(df_2017_com$BMI[df_2017_com$SEX==1])
hist(BMI_men)
hist(df_2017_com$HDLC[df_2017_com$SEX==1])
hist(HDLC_men)
barplot(table(df_2017_com$CURSMOKE[df_2017_com$SEX==1]))
barplot(table(CURSMOKE_men))
barplot(table(df_2017_com$DIABETES[df_2017_com$SEX==1]))
barplot(table(DIABETES_men))
hist(df_2017_com$TOTCHOL[df_2017_com$SEX==1])
hist(TOTCHOL_men)
hist(df_2017_com$SYSBP_UT[df_2017_com$SEX==1])
hist(SYSBP_UT_men)
hist(df_2017_com$SYSBP_T[df_2017_com$SEX==1])
hist(SYSBP_T_men)
sim_brierscore_women <- c()
set.seed(1)

for (i in 1:500){

n_women <- 2205
SYSBP_women <- rnorm(n_women, 123.70, 20.36)
SEX_women <- rep(2, n_women)
AGE_women <- runif(n_women, 18, 80)
BMI_women <- rnorm(n_women, 29.84, 7.96)
BMI_women <- ifelse(BMI_women < 0, 1, BMI_women)
HDLC_women <- rnorm(n_women, 58.10, 15.68)
HDLC_women <- ifelse(HDLC_women < 0, 1, HDLC_women)
CURSMOKE_women <- rbinom(n_women, 1, 0.14)
BPMEDS_women <- rbinom(n_women, 1, 0.29)
SYSBP_UT_women <- ifelse(BPMEDS_women == 0, SYSBP_women, 0)
SYSBP_T_women <- ifelse(BPMEDS_women == 1, SYSBP_women, 0)
```

```
TOTCHOL_women <- rnorm(n_women, 190.51, 41.20)
TOTCHOL_women <- ifelse(TOTCHOL_women < 0, 1, TOTCHOL_women)
DIABETES_women <- rbinom(n_women, 1, 0.12)
CVD_women <- NA
S_women <- 0
sim_df_2017_women <- data.frame(CVD_women, SYSBP_women, SEX_women, AGE_women, BMI_women, HDLC_women, CU
names(sim_df_2017_women) <- c("CVD", "SYSBP", "SEX", "AGE", "BMI", "HDLC", "CURSMOKE", "BPMEDS", "SYSBP
sim_comp_women <- rbind(framingham_df_women, sim_df_2017_women)

#use 70% of dataset as training set and 30% as test set
sim_comp_sample_women <- sample(c(TRUE, FALSE), nrow(sim_comp_women), replace=TRUE, prob=c(0.7,0.3))
sim_train_comp_women  <- sim_comp_women[sim_comp_sample_women, ]
sim_test_comp_women   <- sim_comp_women[!sim_comp_sample_women, ]

mod_sim_train_comp_women <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES,
      data= sim_train_comp_women, family= "binomial")

sim_test_comp_women$predicted_comp_women <- predict(mod_sim_train_comp_women, sim_test_comp_women, type

sim_fit_women <- glm(S ~ log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES, data = sim_test_comp_women, family = "binomial")
sim.pred.prob.women <- predict(sim_fit_women, type = "response")
sim_test_comp_women$inv_odd_weight_women <- (1-sim.pred.prob.women)/sim.pred.prob.women

sim_brierscore_women[i] <- sum(sim_test_comp_women$inv_odd_weight_women[sim_test_comp_women$S==1] * (si
}

hist(log(df_2017_com$SYSBP[df_2017_com$SEX==2]))
hist(SYSBP_women)
hist(df_2017_com$AGE[df_2017_com$SEX==2])
hist(AGE_women)
hist(df_2017_com$BMI[df_2017_com$SEX==2])
hist(BMI_women)
hist(df_2017_com$HDLC[df_2017_com$SEX==2])
hist(HDLC_women)
barplot(table(df_2017_com$CURSMOKE[df_2017_com$SEX==2]))
barplot(table(CURSMOKE_women))
barplot(table(df_2017_com$DIABETES[df_2017_com$SEX==2]))
barplot(table(DIABETES_women))
hist(df_2017_com$TOTCHOL[df_2017_com$SEX==2])
hist(TOTCHOL_women)
hist(df_2017_com$SYSBP_UT[df_2017_com$SEX==2])
hist(SYSBP_UT_women)
hist(df_2017_com$SYSBP_T[df_2017_com$SEX==2])
hist(SYSBP_T_women)
plot(density(log(df_2017_com$SYSBP[df_2017_com$SEX==2])))
plot(density(log(SYSBP_women)))

brier_data <- data.frame(
  Iteration = 1:500,  # Replace with your actual iterations or time points
  BrierScore = sim_brierscore_men # Replace with your actual Brier Scores
)
```

```r
b1 <- ggplot(brier_data, aes(x = Iteration, y = BrierScore)) +
  geom_line(color="pink3") +  # Creates a line plot
  geom_point(color="pink3") +  # Adds points to each data point
  theme_minimal() +  # A minimalistic theme
  labs(
    title = "Brier Score Over Time for Men Model",
    x = "Iteration",
    y = "Brier Score"
  )
brier_data <- data.frame(
  Iteration = 1:500,  # Replace with your actual iterations or time points
  BrierScore = sim_brierscore_women # Replace with your actual Brier Scores
)

b2 <- ggplot(brier_data, aes(x = Iteration, y = BrierScore)) +
  geom_line(color = "skyblue2") +  # Creates a line plot
  geom_point(color = "skyblue2") +  # Adds points to each data point
  theme_minimal() +  # A minimalistic theme
  labs(
    title = "Brier Score Over Time for Women Model",
    x = "Iteration",
    y = "Brier Score"
  )

grid.arrange(b1, b2, ncol=2)
library(kableExtra)
sim_brierscore_men <- mean(sim_brierscore_men)
sim_brierscore_women <- mean(sim_brierscore_women)


metrics_table <- data.frame(
  "Model" = c("Men Model", "Women Model"),
  "Framingham" = c(framingham_bs_men, framingham_bs_women),
  "Framingham + NHANES" = c(brierscore_men, brierscore_women),
  "Framingham + Simulated NHANES" = c(sim_brierscore_men, sim_brierscore_women)
)

# Create and format the table using kable
metrics_table %>%
kable(
      caption = "Brier Scores Stratified by Sex",
      booktabs = TRUE,
      col.names = c("", "Framingham", "Framingham + NHANES", "Framingham + Simulated NHANES")) %>%
  kable_styling(latex_options = c("HOLD_position", "scale_down"))
```