PHP 2550_Project 3

Himashree Chandru

2023-11-27

Abstract

This collaborative research project assesses the applicability of a cardiovascular risk prediction model, originally developed on a specific age group in the Framingham Heart Study, when applied to a more diverse population represented by the NHANES dataset. The study employs logistic regression models and inverse odds weights to enhance transportability, evaluating model performance using Brier scores. Limitations include potential lack of generalizability due to demographic differences between source and target populations, substantial missing data in NHANES, oversimplified assumptions in data simulations, limited covariates in the risk model, and challenges associated with achieving transportability across diverse populations. The study primarily relies on the Brier score, emphasizing the need for a more comprehensive set of evaluation metrics. Addressing these limitations is essential for a nuanced interpretation of the study's findings and enhancing the model's robustness across varied demographic groups.

Introduction

Cardiovascular disease (CVD) remains a leading cause of morbidity and mortality worldwide, necessitating accurate risk prediction models for timely intervention and prevention. In the realm of predictive modeling, the application of risk prediction models across diverse populations is of paramount importance. This collaborative project, conducted in partnership with Dr. Jon Steingrimsson from the Biostatistics Department, focuses on evaluating the performance of a cardiovascular risk prediction model originally developed on a source population when applied to a target population.

The predictive model under scrutiny originates from the Framingham Heart Study, a seminal investigation comprising participants aged 30-62. Historically, such models, while valuable in their original context, have demonstrated challenges when extrapolated to populations with different demographic compositions, as seen in the example of the Framingham ATP-III model's suboptimal generalization to multi-ethnic populations.

The objectives of this project are two-fold:

- Evaluate performance of a cardiovascular risk prediction model (D'Agostino et al., 2008) in a target population underlying NHANES: The evaluation is conducted using the weighting estimator for the Brier score in the target population (Steingrimsson et al., 2022).
- Conduct the same analysis when the target population is simulated: The simulation process involves utilizing summary statistics from the NHANES data to simulate the target population. The same estimator for weighted Brier score in the target population is used to evaluate the performance of the models in the simulated populations.

Data and Preprocessing

The preprocessing involves using data from the Framingham study to obtain a sample of source population data. From the Framingham data, the variables CVD (indicating whether myocardial infarction, fatal coronary heart disease, atherothrombotic infarction, cerebral embolism, intracerebral hemorrhage, or subarachnoid

hemorrhage or fatal cerebrovascular disease occured during followup), TIMECVD (number of days from baseline exam to first CVD event during followup or number of days from baseline to censor date), SEX (participant sex), TOTCHOL (serum total cholesterol in mg/dL), AGE (age at exam), SYSBP (systolic blood pressure in mmHg; mean of last two of three measurements), DIABP (diastolic blood pressure in mmHg; mean of last two of three measurements), CURSMOKE (indicating whether participants are current smokers), DIABETES (indicating whether the participant is diabetic), BPMEDS (whether anti-hypertensive medication was being used at the time of exam), HDLC (high density lipoprotein cholesterol in mg/dL) and BMI (body mass index) are extracted, and any NA's are removed.

Different variables (`SYSBP_T` and `SYSBP_UT`) are created to store the blood pressure values depending on whether or not the individuals were on anti-hypertensive medication. In addition, any observations corresponding to individuals without cardiovascular events within 15 years were removed (removal of censored data). Two different datasets corresponding to males and females was then created.

Then, from the NHANES data for the year 2017-18, variables corresponding to those selected from the Framingham study (above) are similarly selected to create a separate dataset corresponding to a sample of the target population data. The observations are filtered to only include data corresponding to participants belonging to ages 30-62 so that the eligibility criteria for the source population (Framingham study participants) is met.

Summary statistics for the two datasets are calculated stratified by sex, and presented in the tables below.

```
##                       Stratified by SEX
##                        1                 2                 p      test
##   n                       1110              1468
##   CVD (mean (SD))          0.32 (0.47)       0.16 (0.37)    <0.001
##   TIMECVD (mean (SD)) 7226.18 (2402.62) 7952.63 (1830.88) <0.001
##   SEX (mean (SD))          1.00 (0.00)       2.00 (0.00)    <0.001
##   TOTCHOL (mean (SD))  226.34 (41.49)    246.22 (45.91)   <0.001
##   AGE (mean (SD))       60.08 (8.23)      60.62 (8.41)     0.102
##   SYSBP (mean (SD))    138.90 (21.05)    140.02 (23.74)    0.215
##   DIABP (mean (SD))     81.88 (11.41)     80.33 (11.08)    0.001
##   CURSMOKE = 1 (%)        433 (39.0)        451 (30.7)    <0.001
##   DIABETES = 1 (%)         98 ( 8.8)         99 ( 6.7)     0.058
##   BPMEDS = 1 (%)          124 (11.2)        262 (17.8)    <0.001
##   HDLC (mean (SD))      43.58 (13.36)     53.03 (15.69)   <0.001
##   BMI (mean (SD))       26.21 (3.49)      25.55 (4.25)    <0.001
```

```
##                       Stratified by SEX
##                        1              2              p      test
##   n                      1417           1583
##   SYSBP (mean (SD))   126.07 (16.63) 122.46 (18.76) <0.001
##   DIABP (mean (SD))    77.72 (11.12)  73.25 (12.20) <0.001
##   SEX (mean (SD))       1.00 (0.00)    2.00 (0.00)  <0.001
##   AGE (mean (SD))      47.16 (9.97)   46.75 (9.85)   0.251
##   BMI (mean (SD))      30.19 (6.79)   30.77 (8.37)   0.048
##   HDLC (mean (SD))     47.45 (14.54)  57.59 (16.25) <0.001
##   CURSMOKE = 1 (%)       367 (25.9)     270 (17.1)  <0.001
##   BPMEDS = 1 (%)         324 (75.5)     348 (80.0)   0.134
##   TOTCHOL (mean (SD)) 192.86 (40.71) 195.40 (38.95)  0.098
##   DIABETES = 1 (%)       186 (13.1)     174 (11.0)   0.083
```

Methods

NHANES Target Population

First, the data from the target population (NHANES data) has missing values, which we explore.

```
## # A tibble: 13 x 3
##    variable n_miss pct_miss
##    <chr>     <int>    <dbl>
##  1 SYSBP_T    2240  74.7
##  2 SYSBP_UT   2172  72.4
##  3 BPMEDS     2136  71.2
##  4 SYSBP       468  15.6
##  5 DIABP       468  15.6
##  6 HDLC        316  10.5
##  7 TOTCHOL     316  10.5
##  8 BMI         170   5.67
##  9 DIABETES      1   0.0333
## 10 SEQN          0   0
## 11 SEX           0   0
## 12 AGE           0   0
## 13 CURSMOKE      0   0
```

On looking at the missing value summary, we see that there is a large percentage of missing data in `SYSBP_T` and `SYSBP_UT`, which makes multiple imputation risky. But since the models developed on the source population use these variables to make predictions, these variables cannot be excluded from the analysis without the model performance being significantly affected. As a result, complete case analysis will be employed.

Risk Score Model

The model to be fit is a logistic regression model where CVD is the response variable, and the predictors include the logarithmically transformed variables log(HDLC), log(TOTCHOL), log(AGE), log(SYSBP_UT+1), and log(SYSBP_T+1), as well as the binary variables CURSMOKE and DIABETES (D'Agostino et al., 2008).

Using inverse odds weights from the training sets for model building and inverse odds weights from the test sets for model evaluation was found to be ideal for enhancing transportability (Steingrimsson et al., 2022). As a result, a 70:30 train-test split was employed for both the source and target populations, and logistic regression models were used to obtain the odds of belonging to the source population conditional on the covariates (used in the risk score models: HDLC, TOTCHOL, AGE, SYSBP_UT, SYSBP_T, CURSMOKE, and DIABETES) from the training data. These odds were then used to construct inverse odds weights that were used in weighted regression to construct the risk score models. This process was done for men and women separately as their covariate distributions were found to be different in both source and target data during preprocessing.

Next, to evaluate model performance on the target population, we calculate the Brier score for the test set of the target population using the weighted estimator for Brier score in the target population (Steingrimsson et al., 2022):

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

where $S_i = 1$ implies the observation is in the source population,

$D_{test,i}$ implies the observation is in the test set,

$\hat{o}(X_i)$ gives estimates of the inverse-odds weights in the test set,

Y_i is the observed outcome, and

$g_{\hat{\beta}}(X_i)$ is the predicted probability of CVD.

$\hat{o}(X_i)$ is obtained by fitting a logistic regression model where the response variable is `S`, which indicates if the observation is in the source population, conditional on the same covariates as in the model to obtain weights to develop the risk score model above.

The Brier score results are given in Table 1.

Table 1: Weighted Brier Score Estimates in the Target Population

Model	Brier Score
Men	0.0874423
Women	0.0854169

Simulations

The ADEMP framework:

- Aim: To evaluate the impact of simulating data with similar distributions of covariates to the NHANES data on the estimated Brier scores of cardiovascular risk score models.
- Data generating mechanisms: Each of the covariates in the risk score models, `SYSBP`, `AGE`, `HDLC`, `CURSMOKE`, `BPMEDS`, `TOTCHOL` and `DIABETES`, are simulated according to whether they are continuous or categorical, and their summary statistics from the NHANES data. The continuous variables are simulated are normal random variables with the specified mean and standard distribution equal to that in the NHANES data, and the categorical random variables are simulated as binomial distributions with a trial size of 1 and a success probability equal to the proportion of 1's in those variables in the NHANES data. A sample size of 100 was chosen to mimic the number of observations in the test set of the target population for men and women (95 and 114, respectively). The simulations are run separately for men and women due to the difference in covariate distributions, and the evaluation of two separate models for men and women.
- Estimand: Brier score for the cardiovascular risk score models in the simulated target populations.
- Methods: The simulation size was determined by deciding that the Monte Carlo Standard Error of Bias be less than 0.005. By the formula $MonteCarloSE(Bias) = sqrt(Var(BrierScore)/n_{sim})$ and assuming that $SD(BrierScore) <= 0.2$ (verified), the number of simulations is determined to be 1600. In order to calculate the Brier score in the simulated population, a logistic regression model is first fit on all the test data (test data from the source population and the simulated data) to obtain the odds of membership in the source population. Inverse odds weights are then calculated and used to obtain the weighted Brier score estimate for the cardiovascular risk score model in the simulated target population. This is done separately for the male and female data. A seed value of "1234" was used for the simulation of both male and female data to ensure reproducibility.
- Performance Measures: Bias, Empirical Standard Error, and Mean Squared Error are the performance measures collected.

First, we simulate data for men, and the performance measures are displayed in Table 2.

Table 2: Performance Measures for the Simulations Estimating Brier Scores for Men

Measure	Estimate	MCSE
thetamean	0.0581	NA
thetamedian	0.0570	NA
bias	-0.0293	3e-04
empse	0.0117	2e-04
mse	0.0010	0e+00

Next, we simulate data for women, and the performance measures are displayed in Table 3.

Table 3: Performance Measures for the Simulations Estimating Brier Scores for Women

Measure	Estimate	MCSE
thetamean	0.0294	NA
thetamedian	0.0291	NA
bias	-0.0560	2e-04
empse	0.0064	1e-04
mse	0.0032	0e+00

Results

From the Brier Scores in Table 1, it can be seen that the models transport rather well to the NHANES data, possibly due to the inverse odds weights used in the development of the risk score models. From Tables 2 and 3 (simulations for men and women, respectively), it can be seen that the mean Brier Scores are lower than those in Table 1. This indicates that the model performs slightly better in the simulated target populations than in the NHANES target population, which in turn implies that the covariate distributions of the simulated data might not be completely similar to the actual covariate distributions in the NHANES data. The estimator seems to have a higher bias in the case of the model for women compared to that for men. The empirical standard error for both cases is low, indicating low variability in estimates across simulations. The mean squared error is also low in both cases, indicating that the estimates have high accuracy and low variability.

Limitations

Several key limitations in the study should be considered. Firstly, the risk prediction model, originating from the Framingham Heart Study, may lack generalizability to diverse populations, given its specific age range and potential demographic differences. The substantial missing data in the NHANES target population, particularly in variables like SYSBP_T and SYSBP_UT, raises concerns about the robustness of the analysis, as complete case analysis is employed. Additionally, the simplifications made during the simulation process, assuming similar distributions to NHANES data, might oversimplify the complexities of the true data-generating mechanisms. The risk score model itself is limited by the exclusion of potential confounders, potentially impacting the model's completeness. While inverse odds weights are used to enhance transportability, accurately predicting cardiovascular risk across populations with diverse characteristics remains a challenging task. Finally, reliance on the Brier score as the primary evaluation metric, without considering additional metrics, may provide an incomplete picture of the model's performance.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir = "~/Downloads")

#Packages
install.packages("riskCommunicator", repos = "http://cran.us.r-project.org")
install.packages("nhanesA", repos = "http://cran.us.r-project.org")

library(riskCommunicator)
library(tidyverse)
library(tableone)
```

```r
library(nhanesA)
library(naniar)
library(mice)
library(survey)
library(corrplot)
library(rsimsum)
library(knitr)
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)

framingham_df$CURSMOKE <- as.factor(framingham_df$CURSMOKE)
framingham_df$DIABETES <- as.factor(framingham_df$DIABETES)
framingham_df$BPMEDS <- as.factor(framingham_df$BPMEDS)

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)) %>%
  mutate(S = 1)
# 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")
```

```r
# The NHANES data here finds the same covariates among this national survey data

# blood pressure, demographic, bmi, smoking, and hypertension info
bpx_2017 <- nhanes("BPX_J") %>%
  select(SEQN, BPXSY1, BPXDI1) %>%
  rename(SYSBP = BPXSY1, DIABP = BPXDI1)
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 = ifelse(BPQ050A == 1, 1, 0)) %>%
  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") %>%
  filter(AGE < 63 & AGE > 29) #subset of data that meets the eligibility criteria for the Framingham st

df_2017$CURSMOKE <- as.factor(df_2017$CURSMOKE)
df_2017$DIABETES <- as.factor(df_2017$DIABETES)
df_2017$BPMEDS <- as.factor(df_2017$BPMEDS)

CreateTableOne(data = df_2017[-1], strata = c("SEX"))

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)
```

```r
miss_var_summary(df_2017)
set.seed(1234)

#data sampled from the target population
complete_df_2017 <- df_2017[complete.cases(df_2017),] %>%
  mutate(S = 0)

#filtering by sex
complete_df_2017_men <- complete_df_2017 %>%
  filter(SEX == 1)
complete_df_2017_women <- complete_df_2017 %>%
  filter(SEX == 2)

#to create test and train data sets
ignore_target_men <- sample(c(TRUE, FALSE), nrow(complete_df_2017_men), replace = TRUE, prob = c(0.3,0.
ignore_target_women <- sample(c(TRUE, FALSE), nrow(complete_df_2017_women), replace = TRUE, prob = c(0.

complete_df_2017_men_train <- complete_df_2017_men[!ignore_target_men,]
complete_df_2017_men_test <- complete_df_2017_men[ignore_target_men,]

complete_df_2017_women_train <- complete_df_2017_women[!ignore_target_women,]
complete_df_2017_women_test <- complete_df_2017_women[ignore_target_women,]

#data sampled from the source population
#to create test and train data sets
ignore_source_men <- sample(c(TRUE, FALSE), nrow(framingham_df_men), replace = TRUE, prob = c(0.3,0.7))
ignore_source_women <- sample(c(TRUE, FALSE), nrow(framingham_df_women), replace = TRUE, prob = c(0.3,0

framingham_df_men_train <- framingham_df_men[!ignore_source_men,]
framingham_df_men_test <- framingham_df_men[ignore_source_men,]

framingham_df_women_train <- framingham_df_women[!ignore_source_women,]
framingham_df_women_test <- framingham_df_women[ignore_source_women,]

#combining the training data from both the source and target populations to get the inverse odds weight
men_train_full_df <- rbind(framingham_df_men_train[-1], complete_df_2017_men_train[-1])
women_train_full_df <- rbind(framingham_df_women_train[-1], complete_df_2017_women_train[-1])

#models for odds of belonging to the source population

#men
men_source_lo_train <- glm(S ~ HDLC + TOTCHOL + AGE + SYSBP_UT + SYSBP_T + CURSMOKE + DIABETES,
                     data = men_train_full_df, family = "binomial")
men_train_full_df$odds <- (predict(men_source_lo_train, newdata = men_train_full_df, type = "response"))
men_train_full_df$weights <- 1/men_train_full_df$odds #inverse odds weights
men_train_source <- men_train_full_df %>% filter(S == 1) #only source population data to fit model
men_train_source <- cbind(framingham_df_men_train$CVD, men_train_source) %>%
  rename("CVD" = "framingham_df_men_train$CVD")
men_train_source$CVD <- as.integer(men_train_source$CVD)

#women
women_source_lo_train <- glm(S ~ HDLC + TOTCHOL + AGE + SYSBP_UT + SYSBP_T + CURSMOKE + DIABETES,
                     data = women_train_full_df, family = "binomial")
```

```r
women_train_full_df$odds <- (predict(women_source_lo_train, newdata = women_train_full_df, type = "respo
women_train_full_df$weights <- 1/women_train_full_df$odds
women_train_source <- women_train_full_df %>% filter(S == 1)
women_train_source <- cbind(framingham_df_women_train$CVD, women_train_source) %>%
  rename("CVD" = "framingham_df_women_train$CVD")
women_train_source$CVD <- as.integer(women_train_source$CVD)

#fitting risk score models on the training data from the source population
#men
design_men <- svydesign(~1, weights = men_train_source$weights, data = men_train_source)
mod_men_train <- svyglm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES,
      design= design_men, family= "binomial")

#women
design_women <- svydesign(~1, weights = women_train_source$weights, data = women_train_source)
mod_women_train <- svyglm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES,
      design= design_women, family= "binomial")
#combining the testing data from both the source and target populations to get the inverse odds weights
men_test_full_df <- rbind(framingham_df_men_test[-1], complete_df_2017_men_test[-1])
women_test_full_df <- rbind(framingham_df_women_test[-1], complete_df_2017_women_test[-1])

#models for odds of belonging to the source population

#men
men_source_lo_test <- glm(S ~ HDLC + TOTCHOL + AGE + SYSBP_UT + SYSBP_T + CURSMOKE + DIABETES,
                    data = men_test_full_df, family = "binomial")
men_test_full_df$odds <- (predict(men_source_lo_test, newdata = men_test_full_df, type = "response"))/(1
men_test_full_df$weights <- 1/men_test_full_df$odds #inverse odds weights
men_test_source <- men_test_full_df %>% filter(S == 1) #only source population data to fit model
men_test_source <- cbind(framingham_df_men_test$CVD, men_test_source) %>%
  rename("CVD" = "framingham_df_men_test$CVD")
men_test_source$CVD <- as.integer(men_test_source$CVD)

#brier score
men_test_source$pred_probs <- predict(mod_men_train, newdata = men_test_source, type = "response")
men_test_source$brier_num <- (men_test_source$weights)*((men_test_source$CVD - men_test_source$pred_prob

men_brier_score <- sum(men_test_source$brier_num)/nrow(complete_df_2017_men_test) #0.0874

#women
women_source_lo_test <- glm(S ~ HDLC + TOTCHOL + AGE + SYSBP_UT + SYSBP_T + CURSMOKE + DIABETES,
                    data = women_test_full_df, family = "binomial")
women_test_full_df$odds <- (predict(women_source_lo_test, newdata = women_test_full_df, type = "respons
women_test_full_df$weights <- 1/women_test_full_df$odds
women_test_source <- women_test_full_df %>% filter(S == 1)
women_test_source <- cbind(framingham_df_women_test$CVD, women_test_source) %>%
  rename("CVD" = "framingham_df_women_test$CVD")
women_test_source$CVD <- as.integer(women_test_source$CVD)

#brier score
women_test_source$pred_probs <- predict(mod_women_train, newdata = women_test_source, type = "response"
```

```r
women_test_source$brier_num <- (women_test_source$weights)*((women_test_source$CVD - women_test_source$p

women_brier_score <- sum(women_test_source$brier_num)/nrow(complete_df_2017_women_test) #0.0854

#table of brier scores
brier_df <- data.frame(
  Model = c("Men", "Women"),
  Brier_Score = c(men_brier_score, women_brier_score)
)

kable(brier_df, col.names = c("Model", "Brier Score"), caption = "Weighted Brier Score Estimates in the
set.seed(1234)
sim_brier_function_m <- function(sample_size){
  sim_target_male_df <- data.frame(
  SYSBP = rnorm(n = sample_size, mean = 126.07, sd = 16.63),
  AGE = rnorm(n = sample_size, mean = 47.16, sd = 9.97),
  HDLC = rnorm(n = sample_size, mean = 47.45, sd = 14.54),
  CURSMOKE = rbinom(n = sample_size, size = 1, prob = 0.259),
  BPMEDS = rbinom(n = sample_size, size = 1, prob = 0.755),
  TOTCHOL = rnorm(n = sample_size, mean = 192.86, sd = 40.71),
  DIABETES = rbinom(n = sample_size, size = 1, prob = 0.131),
  S = 0
)

sim_target_male_df$CURSMOKE <- as.factor(sim_target_male_df$CURSMOKE)
sim_target_male_df$DIABETES <- as.factor(sim_target_male_df$DIABETES)
sim_target_male_df$BPMEDS <- as.factor(sim_target_male_df$BPMEDS)

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

#combining the testing data from both the source and target populations to get the inverse odds weights
men_test_full_df_sim <- rbind(framingham_df_men_test[,-c(1,2,6,11)], sim_target_male_df)

#models for odds of belonging to the source population

#men
men_source_lo_test_sim <- glm(S ~ HDLC + TOTCHOL + AGE + SYSBP_UT + SYSBP_T + CURSMOKE + DIABETES,
                      data = men_test_full_df_sim, family = "binomial")
men_test_full_df_sim$odds <- (predict(men_source_lo_test_sim, newdata = men_test_full_df_sim, type = "r
men_test_full_df_sim$weights <- 1/men_test_full_df_sim$odds #inverse odds weights
men_test_source_sim <- men_test_full_df_sim %>% filter(S == 1) #only source population data to fit mode
men_test_source_sim <- cbind(framingham_df_men_test$CVD, men_test_source_sim) %>%
  rename("CVD" = "framingham_df_men_test$CVD")

#brier score
men_test_source_sim$pred_probs <- predict(mod_men_train, newdata = men_test_source_sim, type = "respons
men_test_source_sim$brier_num <- (men_test_source_sim$weights)*((men_test_source_sim$CVD - men_test_sou

men_brier_score_sim <- sum(men_test_source_sim$brier_num)/nrow(sim_target_male_df)
```

```r
return(men_brier_score_sim)
}

men_brier_sim_res <- replicate(1600, sim_brier_function_m(100))
output_men_df <- data.frame(dataset = 1:1600, bs = men_brier_sim_res)
simsum_men <- simsum(data = output_men_df, estvarname = "bs", true = men_brier_score)
rsimsum::kable(simsum_men, stats = c("thetamean", "thetamedian", "bias", "empse", "mse"), caption = "Pe
set.seed(1234)
sim_brier_function_f <- function(sample_size){
  sim_target_female_df <- data.frame(
  SYSBP = rnorm(n = sample_size, mean = 122.46, sd = 18.76),
  AGE = rnorm(n = sample_size, mean = 46.75, sd = 9.85),
  HDLC = rnorm(n = sample_size, mean = 57.59, sd = 16.25),
  CURSMOKE = rbinom(n = sample_size, size = 1, prob = 0.171),
  BPMEDS = rbinom(n = sample_size, size = 1, prob = 0.8),
  TOTCHOL = rnorm(n = sample_size, mean = 195.40, sd = 38.95),
  DIABETES = rbinom(n = sample_size, size = 1, prob = 0.11),
  S = 0
)

sim_target_female_df$CURSMOKE <- as.factor(sim_target_female_df$CURSMOKE)
sim_target_female_df$DIABETES <- as.factor(sim_target_female_df$DIABETES)
sim_target_female_df$BPMEDS <- as.factor(sim_target_female_df$BPMEDS)

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

#combining the testing data from both the source and target populations to get the inverse odds weights
women_test_full_df_sim <- rbind(framingham_df_women_test[,-c(1,2,6,11)], sim_target_female_df)

#models for odds of belonging to the source population

#women
women_source_lo_test_sim <- glm(S ~ HDLC + TOTCHOL + AGE + SYSBP_UT + SYSBP_T + CURSMOKE + DIABETES,
                     data = women_test_full_df_sim, family = "binomial")
women_test_full_df_sim$odds <- (predict(women_source_lo_test_sim, newdata = women_test_full_df_sim, typ
women_test_full_df_sim$weights <- 1/women_test_full_df_sim$odds #inverse odds weights
women_test_source_sim <- women_test_full_df_sim %>% filter(S == 1) #only source population data to fit
women_test_source_sim <- cbind(framingham_df_women_test$CVD, women_test_source_sim) %>%
  rename("CVD" = "framingham_df_women_test$CVD")

#brier score
women_test_source_sim$pred_probs <- predict(mod_women_train, newdata = women_test_source_sim, type = "r
women_test_source_sim$brier_num <- (women_test_source_sim$weights)*((women_test_source_sim$CVD - women_

women_brier_score_sim <- sum(women_test_source_sim$brier_num)/nrow(sim_target_female_df)

return(women_brier_score_sim)
}
```

```r
women_brier_sim_res <- replicate(1600, sim_brier_function_f(100))
output_women_df <- data.frame(dataset = 1:1600, bs = women_brier_sim_res)
simsum_women <- simsum(data = output_women_df, estvarname = "bs", true = women_brier_score)
rsimsum::kable(simsum_women, stats = c("thetamean", "thetamedian", "bias", "empse", "mse"), caption = "
```

References

D'Agostino, Ralph B Sr et al. "General cardiovascular risk profile for use in primary care: the Framingham Heart Study." Circulation vol. 117,6 (2008): 743-53. doi:10.1161/CIRCULATIONAHA.107.699579

Steingrimsson, Jon A et al. "Transporting a Prediction Model for Use in a New Target Population." American journal of epidemiology vol. 192,2 (2023): 296-304. doi:10.1093/aje/kwac128