Project Three

Tara Amruthur

2023-11-27

Using Simulations for the Transportability Analysis of a Model

Tara Amruthur

Abstract

Empirical investigations often span two distinct populations—the study population and the target population. The alignment of these populations is challenging, especially when constructing predictive models intended for application in the target population. Discrepancies between them can compromise model efficacy, emphasizing the need for transportability analysis.

Datasets from the Framingham Study and NHANES are leveraged to explore cardiovascular disease (CVD) prediction across diverse populations. The Framingham Study, initiated in 1948, is a longitudinal cohort investigating CVD epidemiology and risk factors. NHANES evaluates the health and nutritional status of the U.S. population, integrating self-reported and comprehensive examination data.

Population disparities between Framingham and NHANES underscore the imprudence of extrapolating model performance across datasets. Transportability analysis becomes imperative to understand model generalizability. The Brier score, a metric for probabilistic predictions, is employed, but adjustments are made to evaluate transportability. Multiple imputation addresses NHANES' missing data, and gender-specific average Brier scores are computed.

The ADEMP framework guides simulation considerations, encompassing aim, data generating mechanism, estimand, methods, and performance measures. Simulation involves diverse parameter variations for gender-specific data generation, aiming to replicate NHANES demographics.

The evaluation of simulated datasets using bias and mean squared error reveals nuanced differences between genders. Optimal covariance values for minimizing bias and MSE differ for men and women. The Brier score comparison and summary statistics show congruence with NHANES for women but slight discrepancies for men, potentially attributed to data variability.

The data generation process incorporates missing variables, potentially contributing to disparities in Brier scores. Computational constraints limit simulations, raising the possibility of unstable estimates. These considerations highlight areas for refining the data generation process and underscore the importance of meticulous transportability analysis in model development and application across diverse populations.

Introduction

In the pursuit of empirical investigations, the focus often extends to two distinct populations: the target population and the study population. While an ideal scenario presupposes their congruence, this alignment is frequently elusive. The construction of predictive models introduces a notable challenge, given that the model is predicated upon the study population but is intended for application to the target population. Discrepancies in characteristics between these populations can substantially compromise the model's efficacy when deployed on the target population. Hence, an evaluation of predictive accuracy alone proves insufficient; the aspect of transportability assumes paramount importance (Steingrimsson et al. 2023).

To underscore the significance of transportability analysis, we draw upon datasets derived from two disparate sources: The Framingham Study and the National Health & Nutrition Examination Survey (NHANES). Initiated by the United States Public Health Service in 1948, the Framingham Study represents an extensive observational cohort investigation dedicated to exploring the epidemiology and risk factors associated with cardiovascular disease (CVD). Evolving into a longitudinal study spanning three generations, this initiative meticulously captures data pertaining to biological and lifestyle risk factors, alongside monitoring cardiovascular, neurological, and other disease outcomes ("Framingham Heart Study," n.d.).

The inaugural cohort of the Framingham Study comprised 5,209 participants aged 28-62, representing two-thirds of the adult populace in Framingham, MA. Subsequently, an offspring cohort was assembled, with a third generation recruited as of 2002. Each participant underwent comprehensive evaluations encompassing medical histories, physician examinations, laboratory tests for vascular risk factors, and, in some instances, cognitive assessments and brain imaging ("Framingham Heart Study," n.d.).

In contrast, the NHANES, administered by the National Center for Health Statistics (NCHS), constitutes a program of studies aimed at evaluating the health and nutritional status of both adults and children in the United States (CDC 2023). A continuous program with evolving focal points to address contemporary needs, NHANES integrates self-reported data on demographic, socioeconomic, dietary, and health-related factors. Furthermore, it encompasses a comprehensive examination entailing medical, dental, and psychological metrics.

The Framingham Study, being a longitudinal exploration enriched with cardiovascular outcome data, facilitates the construction of models predicting an individual's predisposition to cardiovascular disease (CVD). Conversely, the NHANES data lacks information on cardiovascular outcomes. In this context, the pivotal inquiry emerges: To what extent can we prognosticate the likelihood of cardiovascular disease in the NHANES population utilizing a model constructed on the Framingham data?

```
## [1] 2578   14
```

Data

Preceding the commencement of the transportability analysis, an exploratory data analysis was undertaken. The interrelations among numeric columns within the National Health & Nutrition Examination Survey (NHANES) data are illustrated in Figure 1. It is discerned that systolic blood pressure exhibits a pronounced correlation with age, diastolic blood pressure, and body mass index (BMI). Furthermore, a notable correlation is observed between total cholesterol and systolic blood pressure, along with high-density lipoprotein cholesterol (HDLC).

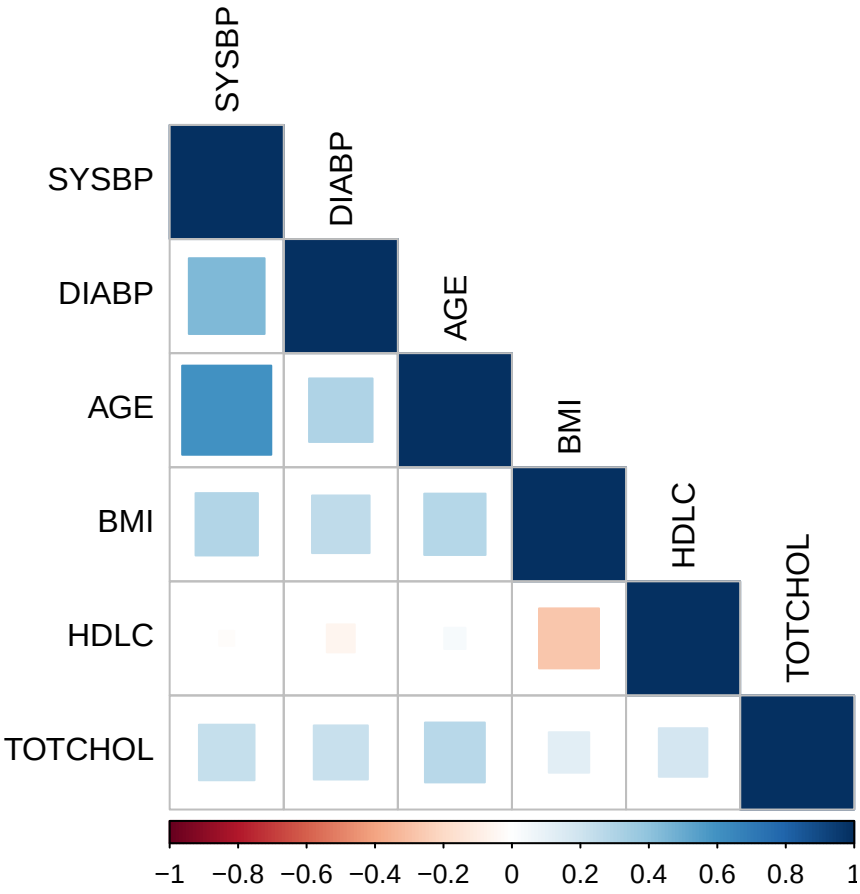


Figure 1. Correlation Plot for NHANES Data

In Figure 2, the inter-column correlations within the numeric variables of the Framingham data are presented. Notably, akin to the NHANES data, discernible correlations are evident between systolic blood pressure and diastolic blood pressure, age, and BMI. However, it is noteworthy that the strength of correlations for other variables in the Framingham data appears comparatively less pronounced than those observed in the NHANES dataset. This discrepancy underscores a substantive distinction between the two datasets.

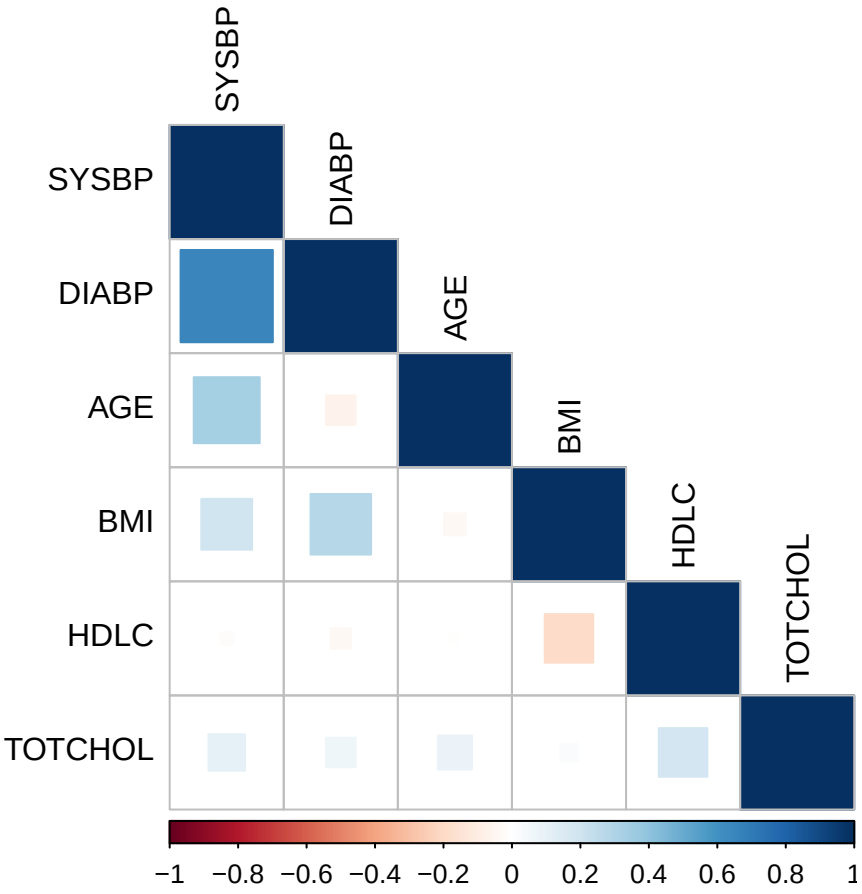


Figure 2. Correlation Plot for Framingham Data

Within the Framingham dataset, an absence of missing data is apparent. Conversely, the NHANES dataset manifests several columns harboring a discernible volume of missing observations, as depicted in Table 1. Notably, a conspicuous pattern emerges wherein the absence of data in the HDLC variable corresponds to missing observations in total cholesterol, and vice versa. Similarly, the absence of systolic blood pressure data coincides with missing data in diastolic blood pressure. Although no other overt patterns are evident, it is imperative to highlight the substantial number of missing observations within the blood pressure medication column.

The ramifications of this observation become evident when considering Table 1; a focus solely on complete cases within the dataset would result in a significant loss of information. Specifically, a mere 1,480 complete observations stand in stark contrast to the 9,254 observations comprising the entire dataset. This underscores the importance of addressing missing data systematically, as reliance solely on complete cases may entail a notable sacrifice of valuable information inherent in the dataset.

Table 1: Number of Missing Values by Column (NHANES)

	Missing Values
SEQN	0
SYSBP	2952
DIABP	2952
SEX	0
AGE	0
BMI	1249
HDLC	2516
CURSMOKE	3398
BPMEDS	7312
TOTCHOL	2516
DIABETES	361

Table 2 presents a comprehensive summary of data pertaining to male participants from both the Framingham and NHANES studies, while Table 3 provides an analogous summary for female participants. Notably, both tables accentuate pivotal distinctions inherent in the cohorts under scrutiny. Specifically, a greater prevalence of individuals in the NHANES dataset is observed to be under blood pressure medication, afflicted with diabetes, or exhibiting elevated body mass index (BMI). In contrast, discernible characteristics within the Framingham population encompass a comparatively advanced mean age and elevated cholesterol levels.

These population disparities assume significance as they elucidate potential factors contributing to divergent model performances across datasets. Specifically, the discerned dissimilarities underscore the imprudence of extrapolating the performance of a model trained on one dataset, exemplified by the Framingham data, to another distinct dataset, such as NHANES. The age-related variations, differences in health conditions, and divergent anthropometric profiles between the populations underscore the necessity for meticulous transportability analysis. Such analysis becomes imperative to glean insights into the generalizability of models derived from one population to another with disparate demographic and clinical characteristics.

Table 2. Summary Statistics for Men by Study

	Study	
Characteristic	**FHS**, N = 1,094[1]	**NHANES**, N = 342[1]
Total Cholesterol	226 (41)	186 (43)
Age	60 (8)	52 (9)
BMI	26.2 (3.5)	31.9 (6.8)
Systolic Blood Pressure	139 (21)	133 (19)
High Density Lipoprotein Cholesterol (HDLC)	44 (13)	46 (15)
Current Smoker	425 (39%)	87 (25%)
Blood Pressure Medication	123 (11%)	262 (77%)
Diabetes	96 (8.8%)	95 (28%)
Cardiovascular Disease	360 (33%)	0 (NA%)
Unknown	0	342
Diastolic Blood Pressure	82 (11)	81 (12)

[1]Mean (SD); n (%)

Table 3. Summary Statistics for Women by Study

Characteristic	Study	
	FHS, N = 1,445[1]	**NHANES**, N = 345[1]
Total Cholesterol	246 (46)	201 (41)
Age	61 (8)	52 (8)
BMI	25.5 (4.2)	34.1 (8.5)
Systolic Blood Pressure	140 (24)	134 (20)
High Density Lipoprotein Cholesterol (HDLC)	53 (16)	56 (17)
Current Smoker	445 (31%)	65 (19%)
Blood Pressure Medication	259 (18%)	271 (79%)
Diabetes	95 (6.6%)	80 (23%)
Cardiovascular Disease	242 (17%)	0 (NA%)
Unknown	0	345
Diastolic Blood Pressure	80 (11)	77 (13)

[1]Mean (SD); n (%)

Transportability Analysis

A metric commonly employed for assessing the precision of probabilistic predictions is the Brier score, particularly pertinent in the context of binary outcomes, such as the occurrence of cardiovascular disease. In this scenario, a predictive model furnishes probabilities for each potential outcome, indicating, for instance, a 10% probability of cardiovascular disease and a 90% probability of its absence. The Brier score is computed as the mean of the squared differences between these predictions and the corresponding actual outcomes (binary values of 0 or 1), expressed by the following formula:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(f_t - o_t)^2$$

where f_t represents the prediction for the tth event and o_t represents the corresponding observed outcome (Goldstein-Greenwood 2021).

However, to evaluate the transportability of our model from the source population to the target population, adjustments are necessary to the Brier score formula. Introducing an indicator variable S, where a value of 1 signifies that an observation originates from the source population and 0 indicates derivation from the target population, leads to the modified Brier score formula:

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

Here, $\hat{o}(X_i)$ represents an estimator for the inverse odds weights, defined as:

$$\hat{o}(X) = \frac{Pr[S = 0|X]}{Pr[S = 1|X]}$$

(Steingrimsson et al. 2023).

Given the substantial amount of missing data within the NHANES dataset, multiple imputation techniques were employed. Five imputed datasets were generated, and for each of these datasets, individuals not meeting the criteria of the Framingham study were filtered out. Subsequently, the Brier score was computed for each filtered dataset, and these values were averaged to yield the gender-specific average Brier scores, as detailed in Table 4.

Table 4: Brier Scores by Gender

	Brier Score
Male	0.0917429
Female	0.0240941

Data Simulation

The ADEMP framework, an invaluable tool for conceptualizing simulation challenges, delineates five pivotal considerations antecedent to the initiation of a simulation study. These factors encompass aim, data generating mechanism, estimand, methods, and performance measures.

Aims: The objective of this endeavor is to employ summary-level data adeptly to simulate individual-level data representative of our target population, constituting the demographic substratum underpinning the NHANES survey data.

Data Generating Mechanism: Data generation was dichotomized by gender, maintaining a uniform procedural methodology with divergent parameter values. Notably, the interdependence between HDLC and cholesterol prompted their simulation through a multivariate normal distribution, allowing for varied associations. Age, characterized by a uniform distribution within the range of 28 to 62, was generated accordingly. Systolic blood pressure followed a normal distribution. The binary variable for smoking was modeled using a binomial distribution, employing probabilities extracted from summary tables. The variable denoting blood pressure medication usage was contingent on varying systolic blood pressure thresholds, ranging from 120 to 140, referencing criteria elucidated by the Mayo Clinic (Clinic 2022). A predictive model for diabetes, predicated on age, HDLC, and blood pressure medication status, was constructed employing NHANES data and subsequently deployed to generate diabetes status for each data point. To circumvent computational constraints, 500 iterations were executed for each permutation of parameter variations. A seed of 1 was used.

Estimand: The estimand, denoted as $\hat{\psi}_{\hat{\beta}}$, assumes the form of the Brier Risk estimator, serving as a metric to gauge model performance within the target population.

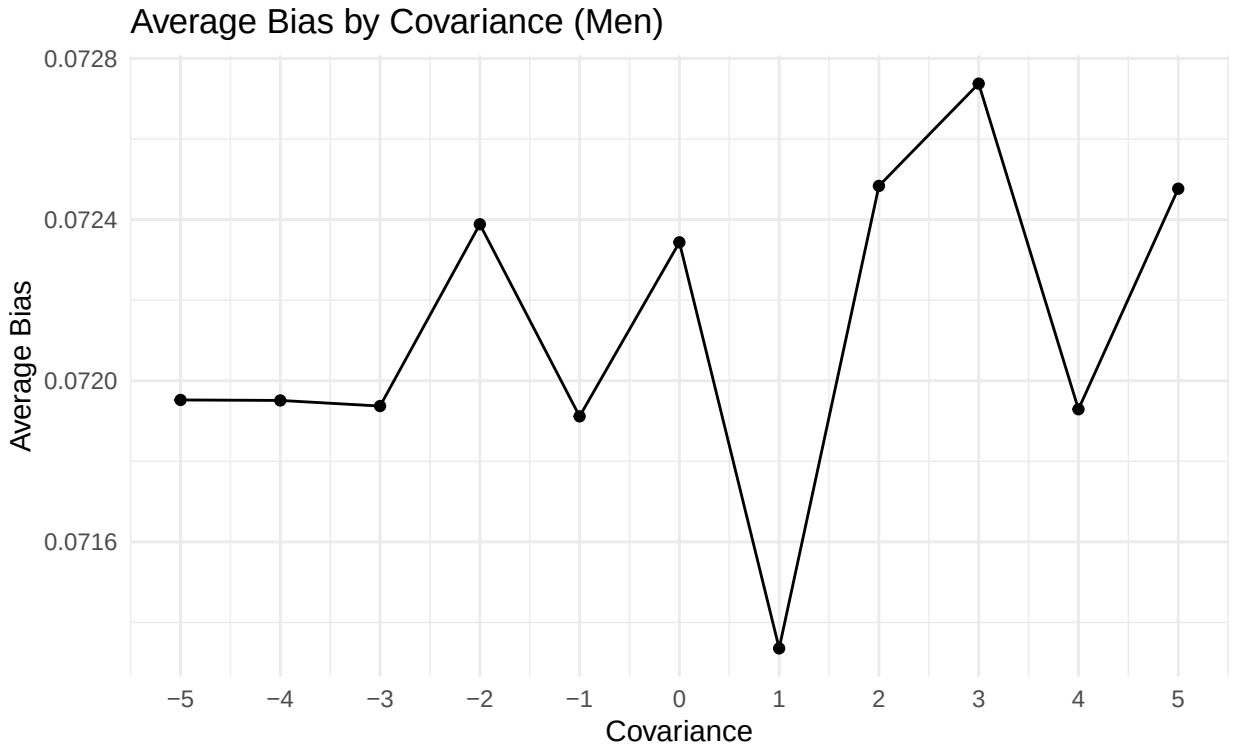
Methods: The estimand was computed using the original NHANES dataset and applied to diverse simulated datasets to discern the variant that most faithfully mirrors the inherent population structure of NHANES.

Performance Measures: Both mean squared error and bias were employed as performance measures, affording a quantitative assessment of the accuracy inherent in the simulated data.

Results

The evaluation of simulated datasets entailed the application of two distinct metrics: bias and mean squared error (MSE). Although the discrepancies in bias across covariance values exhibited a relatively modest amplitude, Figure 3 provides valuable insights into the covariance values, elucidating the relationship between total cholesterol and high-density lipoprotein cholesterol (HDLC), which yield the least bias, on average. Specifically, for the male cohort, the minimum bias is observed at a covariance value of 1, whereas for the female cohort, the minimum bias is observed at a covariance value of 5.

Regarding MSE, analogous marginal variations in bias are observed. Notably, the estimated Brier score in the simulated data demonstrates a more pronounced deviation from the actual Brier score in the male dataset compared to the female dataset. A meticulous examination of Figure 4 reveals that, congruent with the bias findings in Figure 3, the optimal covariance values for minimizing MSE align at 1 for men and 5 for women.


Average Bias by Covariance (Men)
Average Bias
0.0728
0.0724
0.0720
0.0716
−5
−4
−3
−2
−1
0
1
2
3
4
5
Covariance

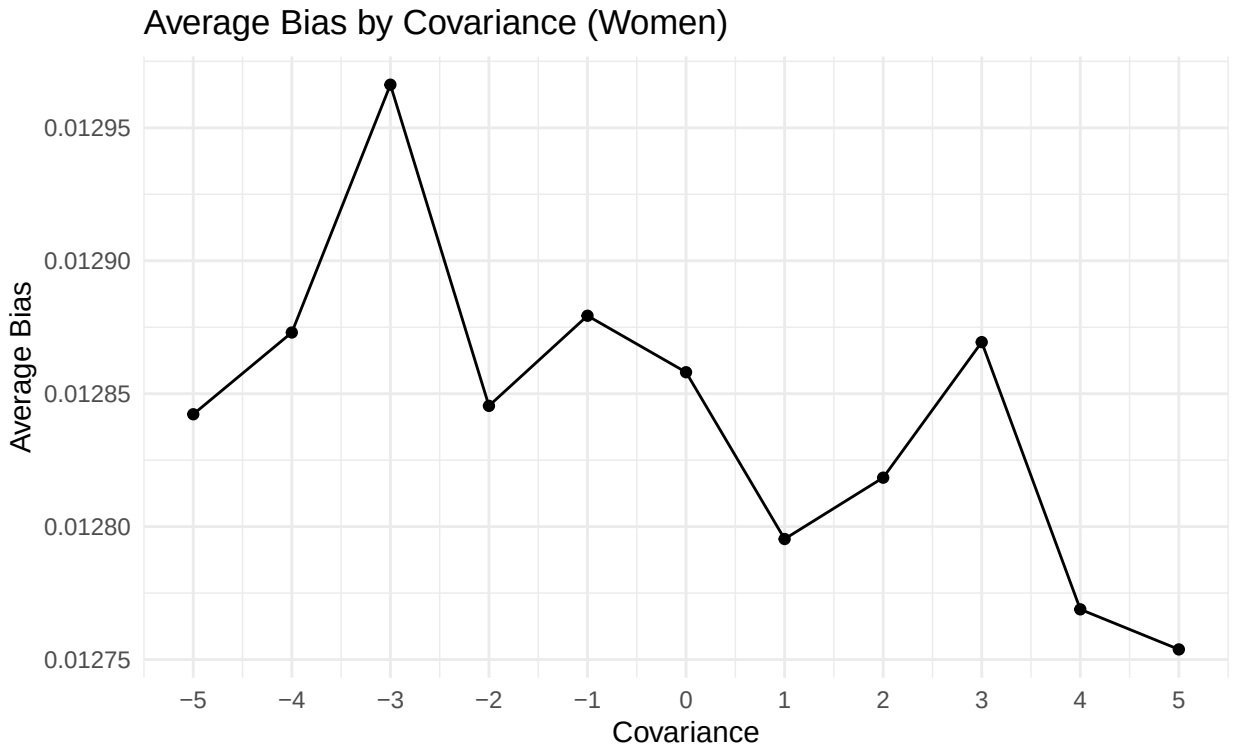
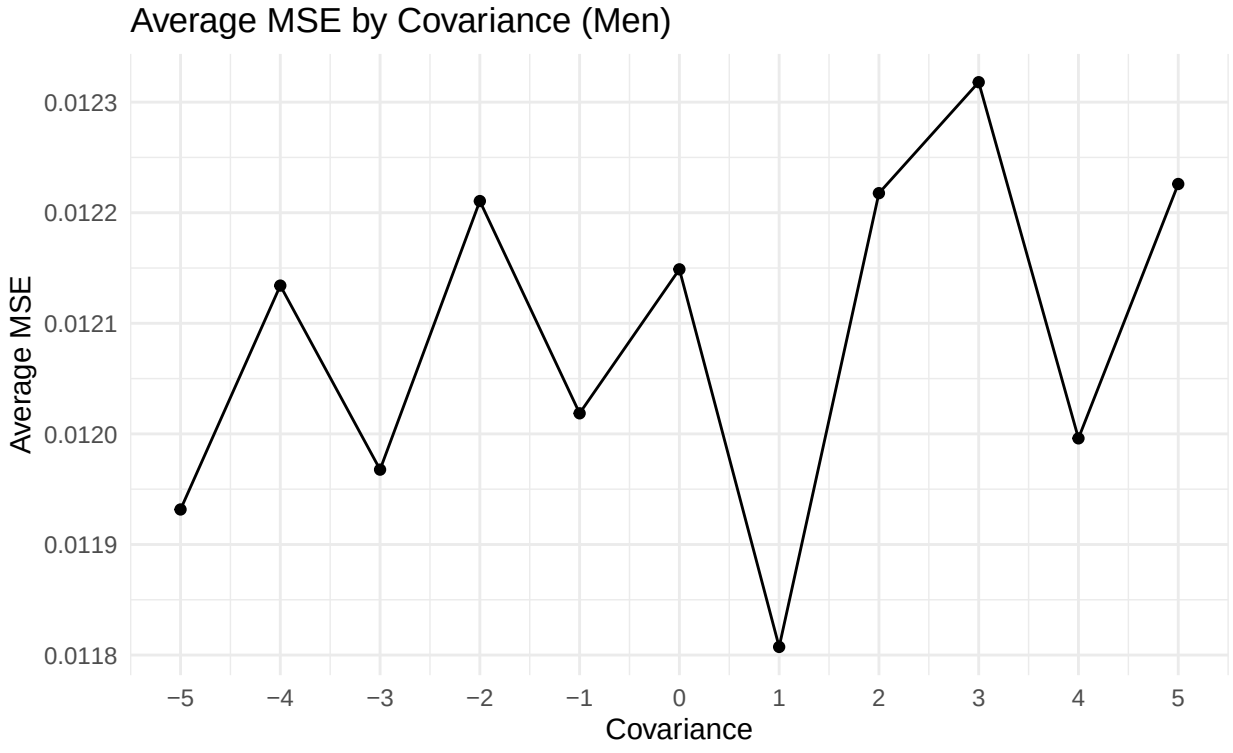

Average Bias by Covariance (Women)
Average Bias
0.01295
0.01290
0.01285
0.01280
0.01275
−5
−4
−3
−2
−1
0
1
2
3
4
5
Covariance

Figure 3. Biases for Men and Women by Covariance Values



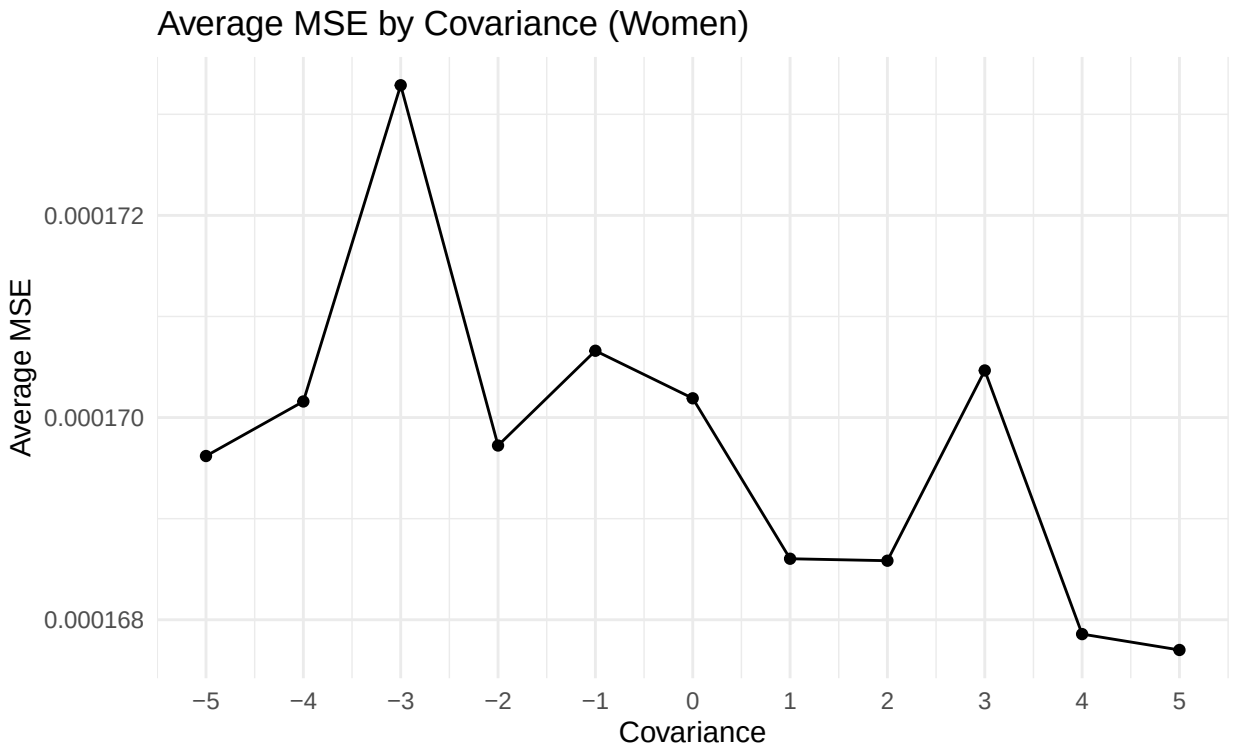


Figure 4. Mean Squared Errors for Men and Women by Covariance Values

Furthermore, we sought to elucidate the influence of varying thresholds in delineating eligibility for blood pressure medication on the absolute bias. This investigation is depicted in Figure 5. The bias, under the

conditions where the covariance was fixed at the optimal values previously identified for each gender, was systematically charted across the spectrum of conceivable thresholds.

Observationally, for the female cohort, the bias showcases minimal fluctuations irrespective of the threshold values. Conversely, a discernible trend emerges indicating that a lower threshold yields a diminished absolute bias, on average, compared to a higher threshold. In contradistinction, the male cohort exhibits a more pronounced sensitivity to the chosen thresholds. Notably, a threshold of 135 results in a near absence of bias, signifying a notable impact on the absolute bias relative to alternate threshold values.

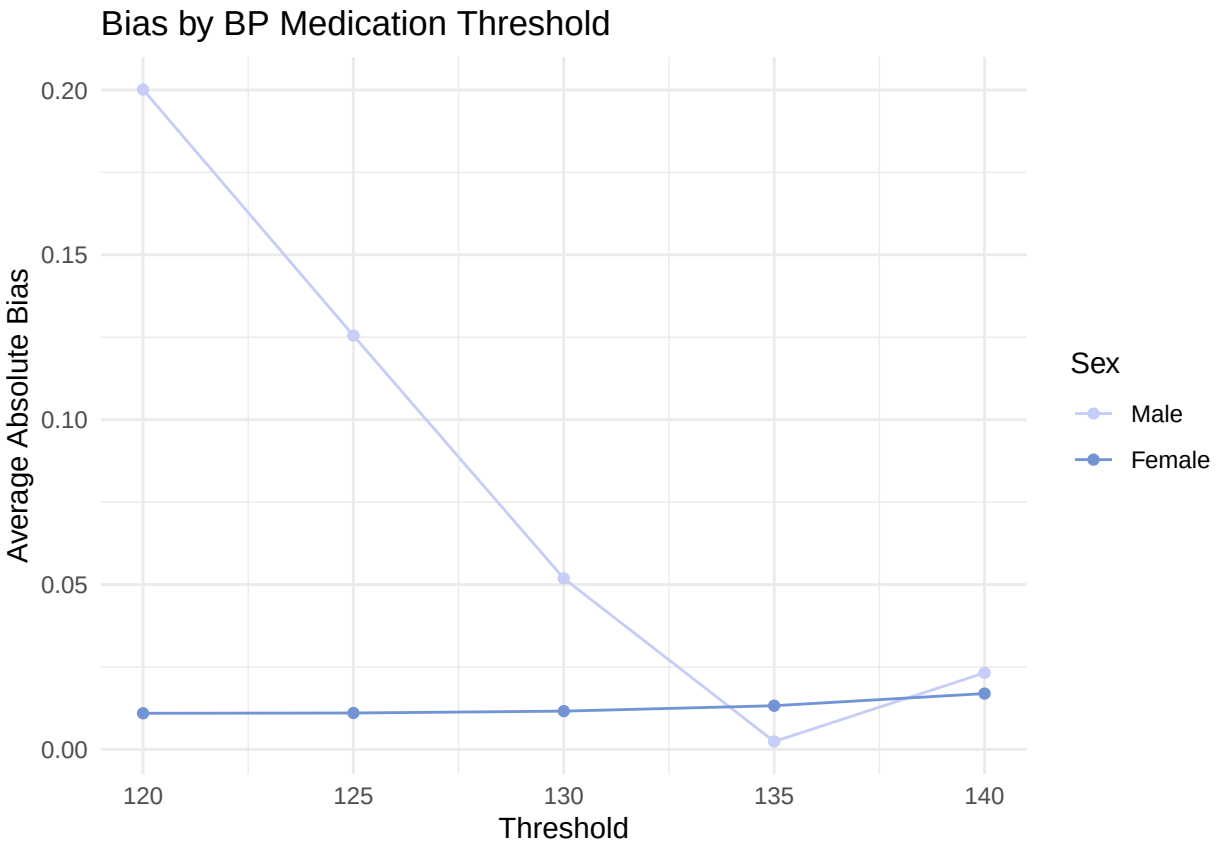


Figure 5. Biases for Men and Women by Threshold Values

Employing the previously established covariance and parameter thresholds, the dataset underwent simulation, and subsequently, the Brier score was computed. Additionally, comprehensive summary tables of the simulated data were generated, facilitating a comparative analysis of its congruence with the summary tables integral to the initial data generation process.

Table 5: Brier Scores of Best Simulated Data

	Brier Score
Male	0.2507895
Female	0.0248542

Table 6. Summary Statistics for Men

Characteristic	**N = 1,500**[1]
Total Cholesterol	186 (43)

Age	45 (10)
Systolic Blood Pressure	117 (20)
High Density Lipoprotein Cholesterol (HDLC)	46 (15)
Current Smoker	418 (28%)
Blood Pressure Medication	293 (20%)
Diabetes	235 (16%)

[1] Mean (SD); n (%)

Table 7. Summary Statistics for Women

Characteristic	**N = 1,680**[1]
Total Cholesterol	200 (42)
Age	45 (10)
Systolic Blood Pressure	134 (20)
High Density Lipoprotein Cholesterol (HDLC)	56 (18)
Current Smoker	323 (19%)
Blood Pressure Medication	1,265 (75%)
Diabetes	286 (17%)

[1] Mean (SD); n (%)

Discussion

As elucidated in Table 5, the Brier score exhibits a closer concordance with the actual Brier score derived from NHANES data for women than for men. Specifically, the Brier score for the simulated male dataset displays an approximate deviation of 0.11. While this discrepancy does not reach a magnitude considered substantial, its origin may be ascribed to the inherent variability within the dataset. Notably, the simulation procedure, relying on mean and standard deviation parameters, might inadequately address the potential influence of outliers or distributional skewness, thus contributing to this anticipated variance.

Upon scrutinizing the summary statistics presented in both Table 6 and Table 7, a salient observation emerges. By employing the multivariate normal distribution for HDLC and total cholesterol, we successfully generated individual-level data in the simulated datasets that closely approximates the actual NHANES data for both genders. Nevertheless, discernible disparities between the simulated and actual datasets manifest, particularly concerning variables such as age, systolic blood pressure, and, notably, blood pressure medication usage within the male dataset.

Limitations & Future Research

These findings prompt a consideration of potential refinements to the data generation process. For instance, the utilization of alternative distributions may enhance the generation of variables such as systolic blood pressure and age. Furthermore, the present data generation strategy encompasses only the variables employed in the Framingham models, yet there may be merit in incorporating additional variables present in the dataset but excluded from the model. Variables such as BMI and diastolic blood pressure, which exhibit relatively robust relationships with other variables according to correlation plots, could be judicious additions to the data generation process to foster a more comprehensive representation of the underlying population.

The data generation process relied on a dataset characterized by the presence of several missing variables. The extent of this missing data may be a contributing factor to the observed disparities in Brier scores between the simulated and actual datasets. Furthermore, the limited number of simulations conducted was constrained by computational and time considerations. This constraint introduces the possibility of unstable Brier score estimates for various parameter combinations.

Conclusion

In conclusion, this study emphasizes the challenge of aligning study and target populations in predictive modeling. Using Framingham and NHANES datasets, we assessed the transportability of cardiovascular disease (CVD) prediction models, revealing population disparities. Despite certain congruences, variations in age, blood pressure, and medication usage suggest areas for model refinement.

The impact of missing data and computational constraints was acknowledged, prompting the exploration of alternative distributions and additional variables for improved simulations. While Brier score discrepancies were observed, not reaching substantial levels, our findings underscore the need for ongoing refinement in predictive modeling methodologies.

Future research should prioritize refining simulation techniques, addressing missing data issues, and exploring variable inclusion strategies to enhance model accuracy. This study highlights the complexity of predictive modeling across diverse populations, emphasizing the ongoing need for improvements in pursuit of reliable and transportable models.

Acknowledgements

This analysis has been performed with help from Dr. Jon Steingrimsson from the School of Public Health at Brown University.

Code Appendix:

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE,
    message = FALSE)
# Load in packages
library(riskCommunicator)
library(tidyverse)
library(tableone)
library(kableExtra)
library(comprehenr)

# Pre-processing for Framingham data
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 %>%
    dplyr::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)) %>%
    dplyr::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 Log transforms - data is
```

```r
# skewed otherwise
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")
library(nhanesA)
# NHANES data pre-processing 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")

# CreateTableOne(data = df_2017, strata =
# c('SEX'))
library(corrplot)
```

```
numeric_nhanes <- df_2017 %>%
    dplyr::select(SYSBP, DIABP, AGE, BMI, HDLC, TOTCHOL)
numeric_nhanes <- numeric_nhanes[complete.cases(numeric_nhanes),
    ]
corrplot(cor(numeric_nhanes), type = "lower", method = "square",
    tl.col = "black")
numeric_fram <- framingham_df %>%
    dplyr::select(SYSBP, DIABP, AGE, BMI, HDLC, TOTCHOL)
numeric_fram <- numeric_fram[complete.cases(numeric_fram),
    ]
corrplot(cor(numeric_fram), type = "lower", method = "square",
    tl.col = "black")
# Patterns of Missingness No missing values in
# Framingham Multiple Imputation
missing_2017 <- as.data.frame(apply(X = is.na(df_2017),
    MARGIN = 2, FUN = sum))
colnames(missing_2017) <- "Missing Values"

kbl(missing_2017, booktabs = T, escape = F, caption = "Number of Missing Values by Column (NHANES)") %>%
    kable_styling(latex_options = "HOLD_position")
library(gtsummary)

complete_cases <- df_2017[complete.cases(df_2017),
    ]
eligible_2017 <- complete_cases %>%
    filter(AGE >= 28 & AGE <= 62) %>%
    dplyr::select(c("SEX", "TOTCHOL", "AGE", "BMI",
        "SYSBP", "HDLC", "CURSMOKE", "BPMEDS", "DIABETES",
        "DIABP"))

eligible_2017$CVD <- NA

eligible_2017$SYSBP_UT <- ifelse(eligible_2017$BPMEDS ==
    0, eligible_2017$SYSBP, 0)
eligible_2017$SYSBP_T <- ifelse(eligible_2017$BPMEDS ==
    1, eligible_2017$SYSBP, 0)

framingham_df <- framingham_df %>%
    dplyr::select(c("SEX", "TOTCHOL", "AGE", "BMI",
        "SYSBP", "HDLC", "CURSMOKE", "BPMEDS", "DIABETES",
        "CVD", "SYSBP_UT", "SYSBP_T", "DIABP"))

final_df <- rbind(framingham_df, eligible_2017)

final_df <- final_df %>%
    mutate(S = case_when(is.na(CVD) ~ 0, TRUE ~ 1))

final_df %>%
    mutate(COHORT = case_when(S == 1 ~ "FHS", S ==
        0 ~ "NHANES")) %>%
    mutate(GENDER = case_when(SEX == 1 ~ "Male", SEX ==
        2 ~ "Female")) %>%
    filter(GENDER == "Male") %>%
```

```
    tbl_summary(by = COHORT, include = (-c(S, GENDER,
        SEX, SYSBP_UT, SYSBP_T)), statistic = list(all_continuous() ~
        "{mean} ({sd})"), label = c(TOTCHOL ~ "Total Cholesterol",
        AGE ~ "Age", BMI ~ "BMI", SYSBP ~ "Systolic Blood Pressure",
        DIABP ~ "Diastolic Blood Pressure", HDLC ~
            "High Density Lipoprotein Cholesterol (HDLC)",
        CURSMOKE ~ "Current Smoker", BPMEDS ~ "Blood Pressure Medication",
        CVD ~ "Cardiovascular Disease", DIABETES ~
            "Diabetes")) %>%
    modify_spanning_header(c("stat_1", "stat_2") ~
        "**Study**") %>%
    as_gt() %>%
    gt::tab_header(title = "Table 2. Summary Statistics for Men by Study")


final_df %>%
    mutate(COHORT = case_when(S == 1 ~ "FHS", S ==
        0 ~ "NHANES")) %>%
    mutate(GENDER = case_when(SEX == 1 ~ "Male", SEX ==
        2 ~ "Female")) %>%
    filter(GENDER == "Female") %>%
    tbl_summary(by = COHORT, include = (-c(S, GENDER,
        SEX, SYSBP_UT, SYSBP_T)), statistic = list(all_continuous() ~
        "{mean} ({sd})"), label = c(TOTCHOL ~ "Total Cholesterol",
        AGE ~ "Age", BMI ~ "BMI", SYSBP ~ "Systolic Blood Pressure",
        DIABP ~ "Diastolic Blood Pressure", HDLC ~
            "High Density Lipoprotein Cholesterol (HDLC)",
        CURSMOKE ~ "Current Smoker", BPMEDS ~ "Blood Pressure Medication",
        CVD ~ "Cardiovascular Disease", DIABETES ~
            "Diabetes")) %>%
    modify_spanning_header(c("stat_1", "stat_2") ~
        "**Study**") %>%
    as_gt() %>%
    gt::tab_header(title = "Table 3. Summary Statistics for Women by Study")
library(mice)
library(kableExtra)

# Multiple Imputation for NHANES data
m <- 5
mice_data_out <- mice(df_2017[, -1], m, pri = F)

calculate_weights <- function(data) {
    #' Function to calculate the weights
    #' @param data, the data used for building the model
    #'
    #' @return vector of weights

    model <- glm(S ~ ., data = data, family = "binomial")
    S1 <- predict(model, type = "response")

    return((1 - S1)/S1)

}
```

```r
brier_score <- function(data, sex) {
    #' Function to calculate the Brier Score by sex
    #' @param data, the data to analyze
    #' @param sex, the sex to calculate the Brier Score for
    #'
    #' @return brier score

    wt <- data %>%
        filter(SEX == sex & S == 1) %>%
        pull(weights)

    nhanes <- data %>%
        filter(SEX == sex & S == 0) %>%
        nrow()

    framingham <- data %>%
        filter(S == 1 & SEX == sex)

    if (sex == 1) {
        pred = predict(mod_men, type = "response")
        return(sum(wt * (framingham$CVD - pred)^2)/nhanes)
    } else {
        pred = predict(mod_women, type = "response")
        return(sum(wt * (framingham$CVD - pred)^2)/nhanes)
    }
}


# Initialize empty vector for male and female
# Brier scores for each imputed dataset
scores_male <- c()
scores_female <- c()

# Loop through each imputed dataset, calculate
# the weights, and then the Brier scores
for (i in 1:m) {

    data <- complete(mice_data_out, i)
    data <- data %>%
        filter(AGE >= 28 & AGE <= 62) %>%
        select(c("SEX", "TOTCHOL", "AGE", "BMI", "SYSBP",
            "HDLC", "CURSMOKE", "BPMEDS", "DIABETES",
            "DIABP"))
    data$CVD <- NA

    data$SYSBP_UT <- ifelse(data$BPMEDS == 0, data$SYSBP,
        0)
    data$SYSBP_T <- ifelse(data$BPMEDS == 1, data$SYSBP,
        0)

    combined_df <- rbind(framingham_df, data) %>%
        mutate(S = case_when(is.na(CVD) ~ 0, TRUE ~
            1))
```

```
    weights <- calculate_weights(combined_df %>%
        select(-CVD))
    combined_df <- cbind(combined_df, weights)

    scores_male <- c(scores_male, brier_score(combined_df,
        1))
    scores_female <- c(scores_female, brier_score(combined_df,
        2))

}

# Create a table with the average Brier scores
# across imputed datasets by gender
brier_df <- data.frame(brier = c(mean(scores_male),
    mean(scores_female)))
rownames(brier_df) <- c("Male", "Female")
colnames(brier_df) <- c("Brier Score")

kbl(brier_df, booktabs = T, escape = F, caption = "Brier Scores by Gender") %>%
    kable_styling(latex_options = "HOLD_position")
library(MASS)
library(gridExtra)

sim_men <- readRDS("sim_results_men.rds")
sim_women <- readRDS("sim_results_women.rds")

est_men <- mean(scores_male)
sim_men_scores <- sim_men %>%
    group_by(cov, thresh) %>%
    summarise(brier = mean(brier), bias = mean(brier -
        est_men), MSE = mean((brier - est_men)^2))


est_women <- mean(scores_female)
sim_women_scores <- sim_women %>%
    group_by(cov, thresh) %>%
    summarise(brier = mean(brier), bias = mean(brier -
        est_women), MSE = mean((brier - est_women)^2))

best_cov_m <- sim_men_scores %>%
    group_by(cov) %>%
    summarise(avg_brier = mean(brier), avg_bias = mean(bias),
        avg_MSE = mean(MSE))

best_cov_w <- sim_women_scores %>%
    group_by(cov) %>%
    summarise(avg_brier = mean(brier), avg_bias = mean(bias),
        avg_MSE = mean(MSE))


p1 <- ggplot(best_cov_m) + geom_line(aes(x = cov, y = avg_bias),
    col = "black") + geom_point(aes(x = cov, y = avg_bias),
    col = "black") + scale_x_continuous(n.breaks = 10) +
```

```
    labs(title = "Average Bias by Covariance (Men)",
        x = "Covariance", y = "Average Bias") + theme_minimal()

p2 <- ggplot(best_cov_w) + geom_line(aes(x = cov, y = avg_bias),
    col = "black") + geom_point(aes(x = cov, y = avg_bias),
    col = "black") + scale_x_continuous(n.breaks = 10) +
    labs(title = "Average Bias by Covariance (Women)",
        x = "Covariance", y = "Average Bias") + theme_minimal()

grid.arrange(p1, p2, nrow = 2, ncol = 1)

p3 <- ggplot(best_cov_m) + geom_line(aes(x = cov, y = avg_MSE),
    col = "black") + geom_point(aes(x = cov, y = avg_MSE),
    col = "black") + scale_x_continuous(n.breaks = 10) +
    labs(title = "Average MSE by Covariance (Men)",
        x = "Covariance", y = "Average MSE") + theme_minimal()

p4 <- ggplot(best_cov_w) + geom_line(aes(x = cov, y = avg_MSE),
    col = "black") + geom_point(aes(x = cov, y = avg_MSE),
    col = "black") + scale_x_continuous(n.breaks = 10) +
    labs(title = "Average MSE by Covariance (Women)",
        x = "Covariance", y = "Average MSE") + theme_minimal()
grid.arrange(p3, p4, nrow = 2, ncol = 1)
library(wesanderson)
best_thresh_m <- sim_men_scores %>%
    filter(cov == 1) %>%
    group_by(thresh) %>%
    summarise(avg_brier = mean(brier), avg_bias = mean(bias),
        avg_MSE = mean(MSE))

best_thresh_w <- sim_women_scores %>%
    filter(cov == 5) %>%
    group_by(thresh) %>%
    summarise(avg_brier = mean(brier), avg_bias = mean(bias),
        avg_MSE = mean(MSE))

best_thresh_m$sex = 1
best_thresh_w$sex = 2

best_thresh <- rbind(best_thresh_m, best_thresh_w)

ggplot(best_thresh) + geom_line(aes(x = thresh, y = abs(avg_bias),
    col = as.factor(sex))) + geom_point(aes(x = thresh,
    y = abs(avg_bias), col = as.factor(sex))) + scale_color_manual(values = c(`1` = wes_palette("GrandBu
    `2` = wes_palette("GrandBudapest2")[4]), labels = c("Male",
    "Female")) + labs(title = "Bias by BP Medication Threshold",
    col = "Sex", x = "Threshold", y = "Average Absolute Bias") +
    theme_minimal()
diabetes_mod <- readRDS("diabetes_mod.rds")

n_m <- 1500
n_w <- 1680
```

```
chol_var_m <- mvrnorm(n_m, mu = c(46, 186), Sigma = matrix(c(15^2,
    1, 1, 43^2), 2, 2))
chol_var_w <- mvrnorm(n_w, mu = c(56, 201), Sigma = matrix(c(17^2,
    5, 5, 41^2), 2, 2))

HDLC_m <- chol_var_m[, 1]
TOTCHOL_m <- chol_var_m[, 2]
HDLC_w <- chol_var_w[, 1]
TOTCHOL_w <- chol_var_w[, 2]

SYSBP_m <- rnorm(n_m, 118, 20)
SYSBP_w <- rnorm(n_w, 134, 20)

AGE_m <- round(runif(n_m, 28, 62))
AGE_w <- round(runif(n_w, 28, 62))

BP_MEDS_m <- as.numeric(SYSBP_m >= 135)
BP_MEDS_w <- as.numeric(SYSBP_w >= 120)

CURSMOKE_m <- rbinom(n_m, 1, 0.25)
CURSMOKE_w <- rbinom(n_w, 1, 0.19)

SEX_m <- rep(1, n_m)
SEX_w <- rep(2, n_w)

data_m <- data.frame(HDLC = HDLC_m, TOTCHOL = TOTCHOL_m,
    SYSBP = SYSBP_m, AGE = AGE_m, BPMEDS = BP_MEDS_m,
    CURSMOKE = CURSMOKE_m, SEX = SEX_m)
data_w <- data.frame(HDLC = HDLC_w, TOTCHOL = TOTCHOL_w,
    SYSBP = SYSBP_w, AGE = AGE_w, BPMEDS = BP_MEDS_w,
    CURSMOKE = CURSMOKE_w, SEX = SEX_w)

x_vars <- model.matrix(CURSMOKE ~ AGE + HDLC + BPMEDS,
    data = data_m)
probs <- x_vars %*% coef(diabetes_mod)
probs <- exp(probs)/(1 + exp(probs))
data_m$DIABETES <- to_vec(for (i in 1:length(probs)) rbinom(1,
    1, probs[i]))

x_vars <- model.matrix(CURSMOKE ~ AGE + HDLC + BPMEDS,
    data = data_w)
probs <- x_vars %*% coef(diabetes_mod)
probs <- exp(probs)/(1 + exp(probs))
data_w$DIABETES <- to_vec(for (i in 1:length(probs)) rbinom(1,
    1, probs[i]))

data <- rbind(data_m, data_w)

data$CVD <- NA

data$SYSBP_UT <- ifelse(data$BPMEDS == 0, data$SYSBP,
    0)
data$SYSBP_T <- ifelse(data$BPMEDS == 1, data$SYSBP,
```

```
    0)

framingham_df <- framingham_df %>%
    dplyr::select("HDLC", "TOTCHOL", "SYSBP", "AGE",
        "BPMEDS", "CURSMOKE", "SEX", "DIABETES", "CVD",
        "SYSBP_UT", "SYSBP_T")

combined_df <- rbind(framingham_df, data) %>%
    mutate(S = case_when(is.na(CVD) ~ 0, TRUE ~ 1))

weights <- calculate_weights(combined_df %>%
    dplyr::select(-CVD))
combined_df <- cbind(combined_df, weights)

score_male <- brier_score(combined_df, 1)
score_female <- brier_score(combined_df, 2)

brier_scores_sim <- data.frame(brier = c(score_male,
    score_female))
colnames(brier_scores_sim) <- c("Brier Score")
rownames(brier_scores_sim) <- c("Male", "Female")
kbl(brier_scores_sim, booktabs = T, escape = F, caption = "Brier Scores of Best Simulated Data") %>%
    kable_styling(latex_options = "HOLD_position")
data_m %>%
    tbl_summary(include = c(TOTCHOL, AGE, SYSBP, HDLC,
        CURSMOKE, BPMEDS, DIABETES), statistic = list(all_continuous() ~
        "{mean} ({sd})"), label = c(TOTCHOL ~ "Total Cholesterol",
        AGE ~ "Age", SYSBP ~ "Systolic Blood Pressure",
        HDLC ~ "High Density Lipoprotein Cholesterol (HDLC)",
        CURSMOKE ~ "Current Smoker", BPMEDS ~ "Blood Pressure Medication",
        DIABETES ~ "Diabetes")) %>%
    as_gt() %>%
    gt::tab_header(title = "Table 6. Summary Statistics for Men")

data_w %>%
    tbl_summary(include = c(TOTCHOL, AGE, SYSBP, HDLC,
        CURSMOKE, BPMEDS, DIABETES), statistic = list(all_continuous() ~
        "{mean} ({sd})"), label = c(TOTCHOL ~ "Total Cholesterol",
        AGE ~ "Age", SYSBP ~ "Systolic Blood Pressure",
        HDLC ~ "High Density Lipoprotein Cholesterol (HDLC)",
        CURSMOKE ~ "Current Smoker", BPMEDS ~ "Blood Pressure Medication",
        DIABETES ~ "Diabetes")) %>%
    as_gt() %>%
    gt::tab_header(title = "Table 7. Summary Statistics for Women")
```

References

CDC. 2023. "NHANES." *Centers for Disease Control and Prevention*. Centers for Disease Control; Prevention. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

Clinic, Mayo. 2022. "High Blood Pressure (Hypertension)." *Mayo Clinic*. Mayo Foundation for Medical Education; Research. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410.

"Framingham Heart Study." n.d. *Boston Medical Center*. https://www.bmc.org/stroke-and-cerebrovascular-center/research/framingham-study#:~:text =The%20Study%20began%20in%201948,the%20town%20of%20Framingham%2

Goldstein-Greenwood, Jacob. 2021. "A Brief on Brier Scores." *A Brief on Brier Scores | UVA Library*. UVA. https://library.virginia.edu/data/articles/a-brief-on-brier-scores.

Steingrimsson, Jon A, Constantine Gatsonis, Bing Li, and Issa J Dahabreh. 2023. "Transporting a Prediction Model for Use in a New Target Population." *American Journal of Epidemiology* 192 (2): 296–304.