PHP 2550: Project 3

Due: December 3rd by 11:59pm

Abstract

This project, in collaboration with Dr. Jon Steingrimsson, focuses on evaluating cardiovascular risk prediction models in a target population different from the one used in their development. Utilizing data from the Framingham Heart Study and the National Health and Nutrition Examination Survey (NHANES), we built separate prediction models for men and women and assessed their generalizability in the NHANES cohort.

Our transportability analysis involved two approaches: a direct application of these models to the NHANES data and a simulation-based estimation using NHANES summary statistics. The non-simulated application yielded Brier Risk scores of 0.0712 for men and 0.0253 for women. The simulated NHANES data resulted in Brier Risks of 0.0776 for men and 0.0238 for women. These results, being fairly aligned, indicate a strong consistency in model performance across actual and simulated target populations. Both findings demonstrate the robustness of the models across different demographic settings and underscore the importance of considering population differences in predictive modeling.

1. Introduction

1.1 Data Overview

Framingham Heart Study Data

The Framingham Heart Study data was collected from a long-term, ongoing cardiovascular cohort study made up of 2,539 observations. This dataset includes a wide range of cardiovascular risk factors such as age, sex, total cholesterol (`TOTCHOL`), systolic blood pressure (`SYSBP`), diastolic blood pressure (`DIABP`), current smoking status (`CURSMOKE`), diabetes status (`DIABETES`), blood pressure medication status (`BPMEDS`), HDL cholesterol (`HDLC`), body mass index (`BMI`), and specific variables like treated systolic blood pressure (`SYSBP_T`) and untreated systolic blood pressure (`SYSBP_UT`). The dataset also includes information on cardiovascular disease outcomes (`CVD`), which is our outcome for building our prediction model.

NHANES Data

The NHANES dataset is a program of studies designed to assess the health and nutritional status of adults and children in the United States made up of 9,254 observations from 2017. This dataset includes similar variables to the Framingham study, such as systolic blood pressure (`SYSBP`), sex (`SEX`), age (`AGE`), `BMI`, HDL cholesterol (`HDLC`), current smoking status (`CURSMOKE`), blood pressure medication status (`BPMEDS`), total cholesterol (`TOTCHOL`), and diabetes status (`DIABETES`). However, unlike the Framingham data, NHANES does not include long-term outcomes for CVD events.

For this project, we focused on aligning and preprocessing these datasets to ensure compatibility and relevance. The NHANES data was filtered to match the eligibility criteria of the Framingham study, including only individuals aged between 30 and 62 years old (dropping 6,415 observations) and individuals who have never had a heart attack or stroke (dropping 67 observations) resulting in a final dataset of 2,772 observations.

1.2 Project Overview

This project is centered around two primary objectives, each focusing on the transportability analysis of cardiovascular risk prediction models. These models, initially developed using the Framingham Heart Study data, are applied to both actual and simulated NHANES data to evaluate their performance across diverse populations.

Transportability Analysis Using NHANES Dataset

The first objective involves a direct application of the Framingham-derived models to the NHANES dataset. This process begins with the creation of a combined dataset, incorporating variables from both Framingham and NHANES, while the outcome variable `CVD` is exclusively derived from the Framingham study. Separate models for men and women are fitted and evaluated. The key metric for assessing model performance in this objective is the Brier Risk Score, which provides a measure of the accuracy of the predictive models when applied to the NHANES target population.

Transportability Analysis Using Simulated NHANES Dataset

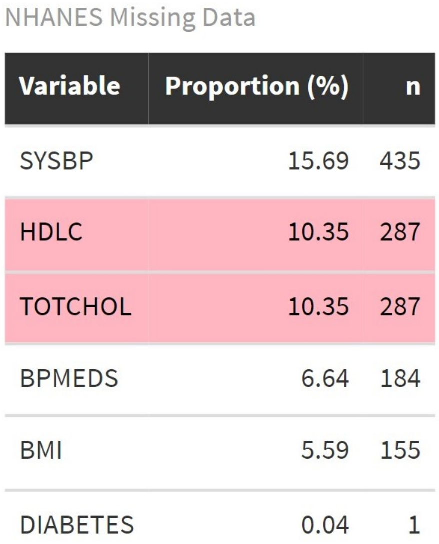
The second objective delves into a more nuanced analysis using the ADEMP framework to simulate the NHANES dataset. The aim here is to juxtapose the transportability analysis results between the actual NHANES data and the simulated counterpart. The data generating mechanism for the simulation draws on the intervariable correlations and distributions from the Framingham dataset, combined with basic descriptive statistics from NHANES. The Brier Risk Score remains the primary estimator for evaluating the transportability of the models. The methods involve a similar process of creating a combined dataset, but this time integrating the Framingham data with the simulated NHANES data. Again, models for each gender are developed and assessed. The performance metric in this objective is the bias observed in the

Brier Risk Scores when comparing the simulated NHANES dataset to the actual NHANES dataset.

2. Methods

2.1 Missing Data

In the course of our analysis, we encountered the challenge of missing data, predominantly within the NHANES dataset. Unlike the Framingham dataset, which presented no missing data issues, NHANES exhibited a significant proportion of missing values across several key variables.



NHANES Missing Data

Variable	Proportion (%)	n
SYSBP	15.69	435
HDLC	10.35	287
TOTCHOL	10.35	287
BPMEDS	6.64	184
BMI	5.59	155
DIABETES	0.04	1

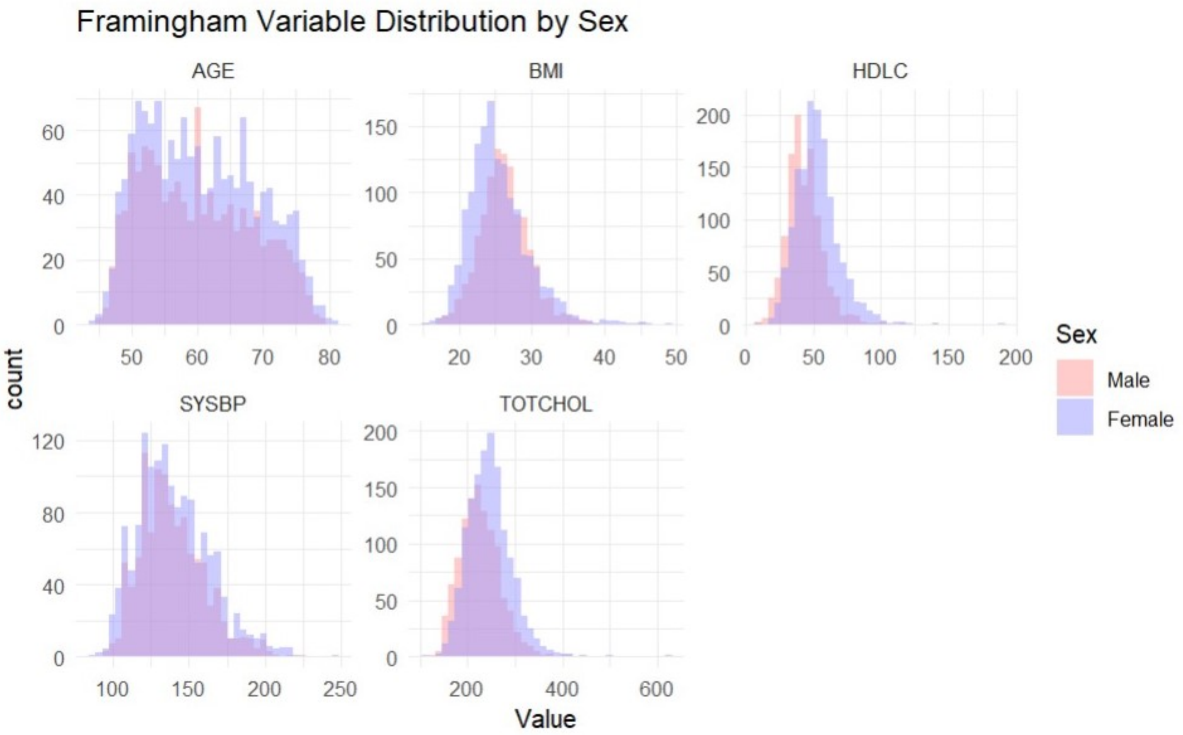
Several variables, including `HDLC` and `TOTCHOL`, exhibited simultaneous missingness, likely caused by missing medical tests. To tackle this issue, multiple imputation was used, this involved the creation of five distinct imputed datasets. For each dataset, we calculated the non-simulated Brier risk score estimates. These scores were then averaged to derive the estimated Brier score, which was used to assess transportability.

2.2 Exploratory Data Analysis

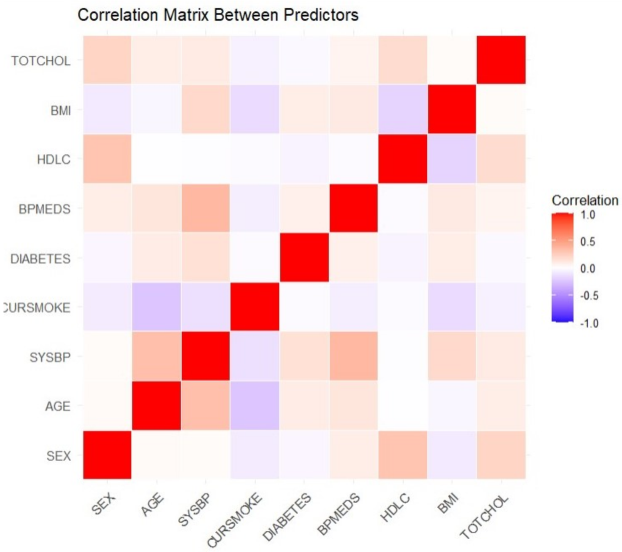
Exploratory data analysis began with a deep dive into the Framingham dataset, which was used in developing our CVD prediction models as well as crucial to creating our simulated NHANES data. A descriptive statistical analysis, stratified by sex, revealed notable differences across several key variables. Here we present the summarized statistics:

Table 1: Framingham Descriptive Statistics by Sex

Variable	Male (Mean (SD))	Female (Mean (SD))	p-value
n	1110	1468	
CVD	0.32 (0.47)	0.16 (0.37)	<0.001
TIMECVD	7226.18 (2402.62)	7952.63 (1830.88)	<0.001
SEX	1.00 (0.00)	2.00 (0.00)	<0.001
TOTCHOL	226.34 (41.49)	246.22 (45.91)	<0.001
AGE	60.08 (8.23)	60.62 (8.41)	0.102
SYSBP	138.90 (21.05)	140.02 (23.74)	0.215
DIABP	81.88 (11.41)	80.33 (11.08)	0.001
CURSMOKE	0.39 (0.49)	0.31 (0.46)	<0.001
DIABETES	0.09 (0.28)	0.07 (0.25)	0.049
BPMEDS	0.11 (0.32)	0.18 (0.38)	<0.001
HDLC	43.58 (13.36)	53.03 (15.69)	<0.001
BMI	26.21 (3.49)	25.55 (4.25)	<0.001



Using the descriptive statistics shown in Table 1 and the variable distributions in the figure above we can see several key details in Framingham's predictor variables. To start the difference in mean values of CVD, HDLC, and BMI, alongside the significant p-values, highlight distinct cardiovascular health profiles between the sexes. We also observed that BMI, TOTCHOL, HDLC, and SYSBP are all skewed to the right, indicating that normal distribution assumptions would not be appropriate for data generation during our simulation study. Moreover, the distributions for HDLC and BMI differed between sexes, suggesting that separating data by sex in the analysis would be neccessary. Lastly, the AGE variable displayed a unique distribution with fewer occurrences near the ages of 40 and 80 and a nearly uniform distribution with slight right skewness between these ages.



In addition to this, to understand the interrelationships among predictor variables a correlation matrix shown above was generated. This revealed various degrees of correlation between our predictor variables such as AGE and SYSBP which had a strong positive correlation between them. This is important as during our data generating process we will want our simulated data to follow a simple structure, under the assumption that the NHANES dataset will exhibit similar intercorrelation relationships as the Framingham dataset.

2.3 Data Generation

The data generation process for simulating the NHANES data began with utilizing the correlation matrix from the Framingham study's predictor variables (excluding sex) to generate a dataset from a multivariate normal distribution using the `mvnorm` function. The resulting values were then transformed into percentiles through the `pnorm` function. This step was essential, as our exploratory data analysis indicated skewed distributions for several Framingham variables, necessitating an approach that would account for this skewness during simulation.

With the generated percentiles in hand, we proceeded to sample from the standardized distributions of the continuous Framingham variables to maintain the skewed and unique distributions. Subsequently, `SEX` was assigned to each observation, with 1600 allocated as females and 1400 as males, approximating the sex distribution observed in NHANES. This approximation was made as at the end of our data generation process simulated individuals that fell outside of the Framingham eligibility age requirements had to be dropped.

This gender assignment was integral in mirroring the true NHANES population as in exploratory analysis of the Framingham data and from the basic descriptive statistics we have available for NHANES we noticed a strong difference between the sexes. The next phase leveraged the NHANES descriptive statistics, to refine our simulated data. The NHANES statistics provided a benchmark for recentering and rescaling the continuous variables from the standardized Framingham distributions:

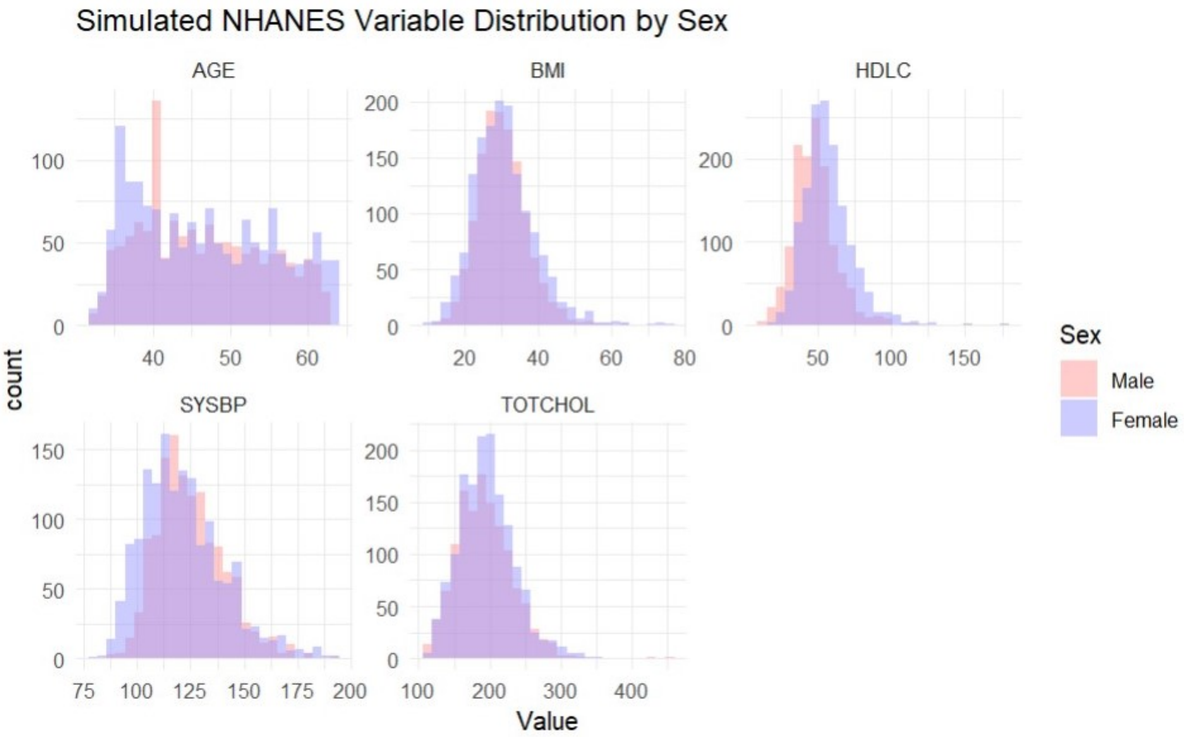
Table 2: NHANES Descriptive Statistics by Sex

Variable	Male (Mean (SD))	Women (Mean (SD))	p-value
n	1300	1462	
AGE	46.88 (9.37)	46.61 (9.25)	0.458
SYSBP	125.77 (16.48)	122.58 (18.92)	<0.001
CURSMOKE	0.26 (0.44)	0.17 (0.38)	<0.001
DIABETES	0.13 (0.33)	0.10 (0.31)	0.068
BPMEDS	0.24 (0.43)	0.22 (0.42)	0.245
HDLC	47.49 (14.44)	57.41 (16.26)	<0.001
BMI	30.25 (6.80)	30.91 (8.45)	0.027
TOTCHOL	193.70 (40.16)	195.98 (38.42)	0.149

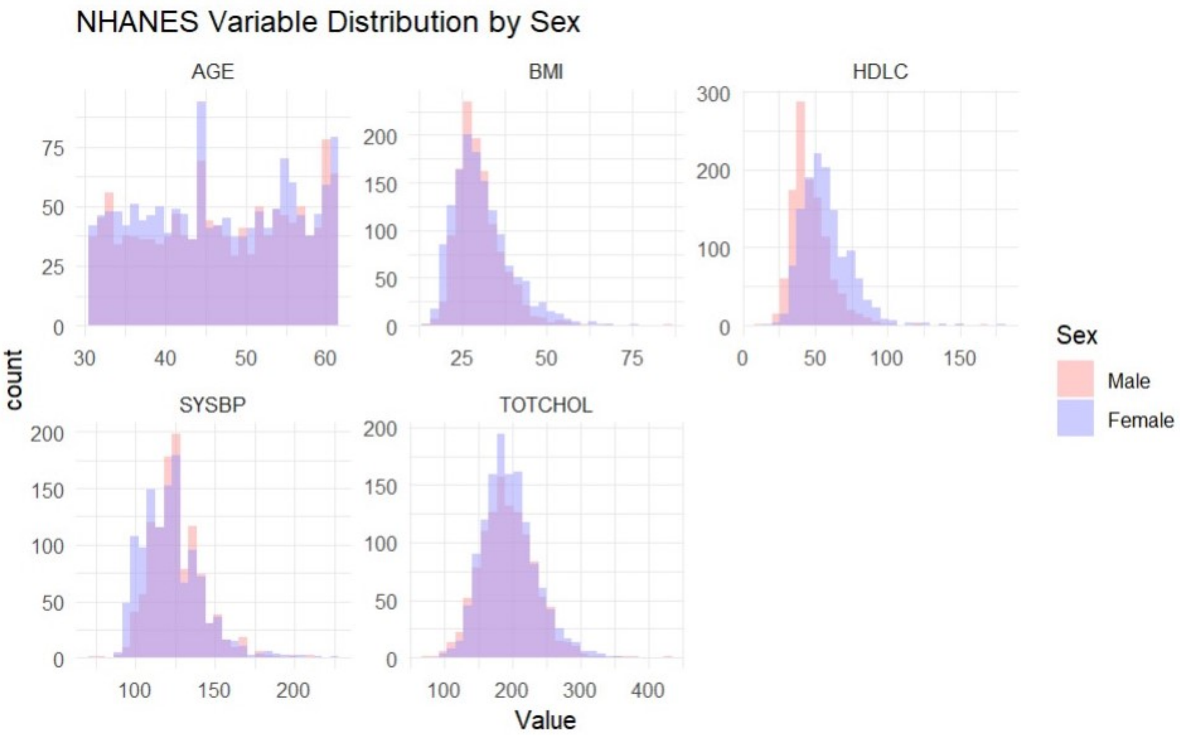
Binary variables such as CURSMOKE, BPMEDS, and DIABETES were converted from norm percentiles using `qbinom`, with thresholds determined by the proportions in NHANES. For the continuous variables, after sampling from the standardized Framingham distributions, they were recentered and rescaled to align with their NHANES counterparts' means and standard deviations. Below shows the descriptive statistics and distributions of one such simulated dataset:

Table 3: Simulated Dataset Descriptive Statistics by Sex

Variable	Sex 1 (Mean (SD))	Sex 2 (Mean (SD))	p-value
n	1274	1509	
AGE	46.62 (8.15)	46.84 (8.70)	0.496
SYSBP	125.01 (15.81)	121.54 (19.17)	<0.001
CURSMOKE	0.27 (0.44)	0.18 (0.38)	<0.001
DIABETES	0.13 (0.34)	0.10 (0.30)	0.020
BPMEDS	0.23 (0.42)	0.20 (0.40)	0.065
HDLC	47.83 (15.07)	56.74 (16.40)	<0.001
BMI	30.16 (6.59)	30.66 (8.38)	0.079
TOTCHOL	193.35 (39.25)	196.17 (37.89)	0.054



From the table above we can see that the means and standard deviations of the simulated data closely resemble those of the actual NHANES values. The figure above also confirms that the simulated dataset distributions also align closely with those from the Framingham variables. Comparing these to the true NHANES distributions:



The true NHANES variable distributions are notably similar to the simulated distributions, with the one major exception of `AGE`, which in NHANES exhibits a more uniform distribution, unlike the Framingham data. Overall, the simulation process effectively replicates the NHANES means and standard deviations while preserving the distributional characteristics observed in the Framingham study, albeit with minor discrepancies in comparison to the true NHANES distributions.

2.4 Brier Risk Estimation

The estimation of Brier risk scores for both simulated and non-simulated NHANES datasets involved a similar methodology. The initial step was to merge the NHANES dataset, whether simulated or actual, with the Framingham dataset. After this, we performed an 80-20% train-test split. Using the training data, we then fit two logistic regression models for each sex to predict the occurrence of cardiovascular disease (CVD). The model used was predefined for this project and took the following form: `glm(CVD ~ log(HDLC) + log(TOTCHOL) + log(AGE) + log(SYSBP_UT + 1) + log(SYSBP_T + 1) + CURSMOKE + DIABETES, family = "binomial")`

Following the model fitting, we proceeded to calculate the estimated Brier score using the formula:

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

Here, S indicates whether the individual is from the Framingham study $S = 1$ or the NHANES dataset $S = 0$, and D differentiates between the training $D = 0$ and test $D = 1$ sets. The function $g(X)$ represents the predictive model for the probability of $Y = 1$ given X in the training set. The term $\hat{o}(X)$ is the estimator for the inverse odds weights in the test set, defined as:

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

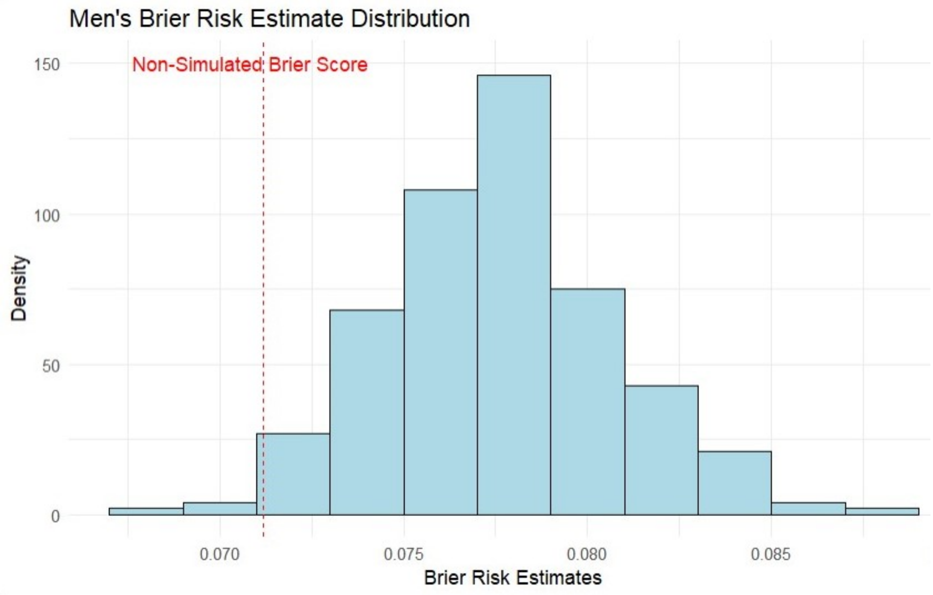
To calculate the inverse-odds weights, we fit a logistic regression model to estimate the probability of an observation coming from the NHANES study as opposed to the Framingham study: `glm(STUDY ~ SEX + AGE + SYSBP + CURSMOKE + DIABETES + BPMEDS + HDLC + BMI + TOTCHOL, family = binomial())`. This model allowed us to estimate the weights needed for the Brier score calculation. With the logistic regression models for both men and women defined, we then applied these models to their respective test datasets to obtain predictions. Combining these predictions with the calculated weights, we could estimate the Brier risk score for each subset.

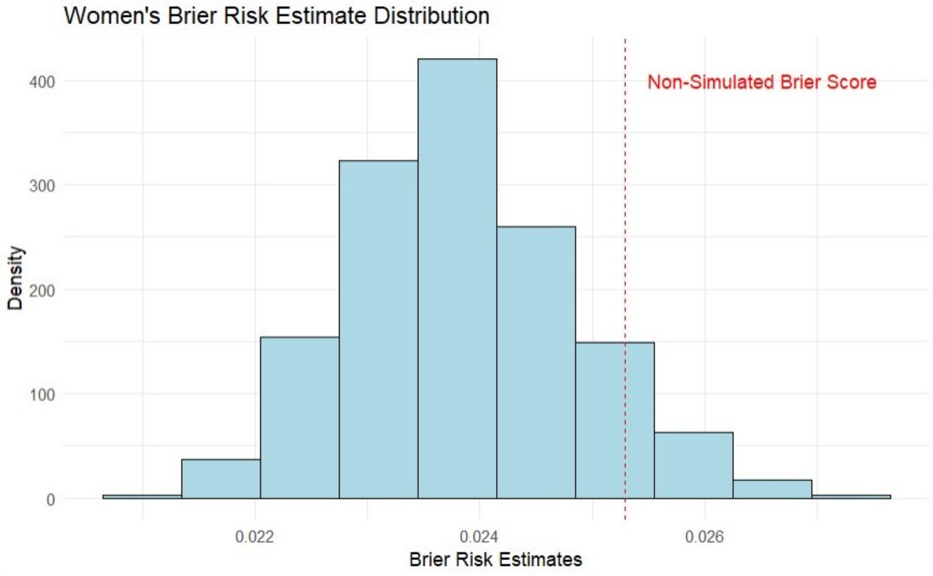
For each of the five imputed datasets, we replicated this process to calculate each of the Brier Risk scores, subsequently averaging these scores to find the score connected to the actual NHANES dataset. In our simulation study, we extended this methodology across 500 simulations. The choice of this number was strategically based on achieving a Standard Error (SE) of the Bias of less than 0.001. We believed that this threshold would offer sufficient precision to accurately gauge the model's bias when using the simulated data. This approach ensured a robust evaluation of the model's performance against the NHANES dataset.

3. Results

The analysis of the non-simulated NHANES dataset provided insightful results regarding the transportability of the CVD prediction models. For the male model, a Brier risk score of 0.0712 was achieved, while the female model demonstrated a Brier risk score of 0.0253. These scores suggest that both models exhibit a high level of transportability to the NHANES data, with the female model performing slightly better. This differential performance may be attributed to various factors, including the nature of the risk factors and their prevalence in each sex or the accuracy of the model in capturing the underlying risk in the population.

Turning our attention to the simulation study, the mean Brier risk scores for the male and female models across all 500 simulations were 0.0776 and 0.0238, respectively. The observed biases were +0.006 for the male model and -0.0015 for the female model relative to the non-simulated scores. These biases, while present, are minimal, indicating that the simulated datasets closely approximated the actual NHANES dataset in terms of the predictive transportability of the CVD models.





The distribution of the estimated Brier risk scores, depicted in the figures above for the male and female model, further corroborates these findings. The distributions suggests a decently strong agreement between the simulated and actual Brier risk scores, reinforcing the notion that the simulated data provides a reliable basis for transportability analysis.

4. Limitations

The simulation study, while yielding promising results, is not without its limitations. A significant constraint stems from the underlying assumptions made during the data generation process. Specifically, we assumed that the NHANES dataset would exhibit intervariable correlations and distributions similar to those found in the Framingham dataset. Although this assumption held reasonably well for most variables, it did not accurately reflect the distribution of ages within the NHANES population, indicating a discrepancy between the simulated and actual age distributions.

This difference in age distribution, albeit minor, highlights a potential pitfall of the simulation approach: the validity of our generated synthetic data is contingent upon the extent to which the source and target populations share similar properties. In scenarios where the source study data and the target population differ significantly, the assumption of comparable distributions and correlations may lead to biased or unrealistic simulations. Consequently, this could impact the generalizability and transportability of the findings derived from such simulated data. Moreover, the Framingham cohort may not fully encapsulate the diversity of the broader population represented in NHANES, particularly in terms of ethnicity and socioeconomic factors.

5. Conclusion

Our study has successfully demonstrated the transportability of cardiovascular disease (CVD) prediction models from the Framingham study to the NHANES population. The non-simulated NHANES data yielded Brier risk scores that show the models' high level of transportability, with the female model exhibiting a marginally superior performance. The simulation study supported these findings, with the mean Brier risk scores from the simulations presenting minimal biases against the true NHANES population, suggesting that the synthetic data approximates the actual NHANES dataset well in terms of generalizability.

However, the simulation study's reliability has several limitations stemming from assumptions in the data generation process. The discrepancy observed in the age distributions between the simulated and NHANES datasets illustrates the caution needed when inferring population-level conclusions from simulated data. The potential for bias introduced by differences between the source and target populations' characteristics highlight the importance of the consideration of demographic diversity in study samples. However despite these limitations, the study shows the utility of transportability analysis using simulation.

Appendix

Github: https://github.com/Kdossal/2550_Project3

Code: https://github.com/Kdossal/2550_Project3/Code