Comparative Analysis of Simulated vs. Non-Simulated Transportability

Caiwei Xiong

Abstract

Background: Risk prediction models, like the Framingham ATP-III model for cardiovascular disease, play a pivotal role in clinical decision-making. However, their effectiveness can be compromised when applied to populations with distinct characteristics. To address this, transportability analysis techniques have been developed to fine-tune model performance to account for demographic differences. Yet, these techniques often depend on individual-level data from the target population, which may not always be accessible. In this context, simulating target population data using summary statistics emerges as a potential solution to this challenge.

Methods: Using summary statistics from the National Health and Nutrition Examination Survey (NHANES) and a range of correlation parameters, we generated data for a simulated target population. We then conducted transportability analyses on each dataset to calculate the model's Brier score and Area Under the Curve (AUC). For each correlation scenario, we compared the bias of these estimates against the Brier score and AUC derived from individual-level NHANES data.

Results: In our transportability analyses, which utilized simulated datasets, we observed relative biases ranging from -0.05 to 1.5 when compared to estimates obtained from individual-level data. Adjustments in the degree of association between specific simulated variables indicated potential for performance improvement. However, it's noteworthy that simulations assuming no associations did not show substantial differences. This finding suggests that the assumed inter-variable associations may have a limited impact on the accuracy of our transportability analyses.

Conclusions: The observation of low relative biases in our study supports the validity of using simulated data for transportability analysis, particularly in estimating Brier scores and Area Under the Curve (AUC) for a target population when individual-level data is unavailable. Furthermore, our findings indicate that researchers can simplify the process of simulating target data by assuming no association between covariates, without significantly compromising the accuracy of the results. This approach could offer a more straightforward and accessible method for conducting such analyses.

Introduction

Transportability analysis plays a pivotal role in epidemiological and statistical research, particularly when examining the applicability of findings from one population to another. This analysis is crucial in understanding how a model, developed in a specific setting or based on a particular dataset, performs when applied to different populations. Our study embarks on this exploration by comparing outcomes between simulated and non-simulated datasets. The simulated dataset, crafted through sophisticated statistical techniques, aims to mirror the characteristics of a real-world population, providing an invaluable tool for assessing model robustness and transportability without compromising individual privacy. On the other hand, the non-simulated dataset, derived from actual population data, offers insights grounded in real-world scenarios. By juxtaposing these two datasets in a transportability analysis framework, we aim to uncover the strengths and limitations of predictive models across varied contexts, enhancing our understanding of their generalizability and utility in diverse epidemiological landscapes.

Motivation datasets

The `Framingham` dataset, accessible through the `RiskCommunicator` package, originates from the renowned Framingham Heart Study. This dataset is pivotal in cardiovascular research, offering extensive data on risk factors associated with heart disease. It includes a variety of variables such as cholesterol levels, blood pressure, diabetes status, smoking habits, and other relevant health indicators. The dataset's longitudinal design, tracking participants over an extended period, allows for in-depth analysis of cardiovascular disease progression and risk factor evolution. The Framingham study's contribution through this dataset has been instrumental in shaping our understanding of heart disease, influencing preventive measures and treatment strategies worldwide. The dataset's comprehensive nature and historical significance make it a cornerstone in the field of cardiovascular research and risk assessment. (Grembi 2022)

The `NHANES` (National Health and Nutrition Examination Survey) package, offers a comprehensive collection of data pertaining to blood pressure measurements. This dataset is instrumental in understanding various health aspects related to medical measurements across a diverse population. It encompasses a wide range of variables, including systolic and diastolic blood pressure readings, which are crucial for medical research and public health analysis. The data, derived from a nationally representative sample, provides insights into the prevalence and distribution of blood pressure-related health conditions in the U.S. population. The NHANES survey's rigorous methodology ensures the reliability and accuracy of dataset, making it a valuable resource for epidemiologists, public health officials, and researchers in related fields. (Endres 2023) To construct a comprehensive dataset from the `NHANES` package that includes information on blood pressure, demographics, BMI, smoking status, and hypertension medication, we employed a multi-step process using R. We began by extracting systolic and diastolic blood pressure measurements from the "`BPX_J`" dataset, assigning new variable names `SYSBP` and `DIABP` for clarity. Demographic details such as sex and age were then retrieved from the "`DEMO_J`" dataset and labeled accordingly. BMI values were sourced from the "`BMX_J`" dataset. For smoking status, the "`SMQ_J`" dataset was queried, creating a binary indicator for current smoking status based on the responses. Similarly, hypertension medication usage was determined from the "`BPQ_J`" dataset, with conditions applied to assign a binary status or an NA for missing data. Total cholesterol levels and HDL cholesterol data were selected from the "`TCHOL_J`" and "`HDL_J`" datasets and were renamed to `TOTCHOL` and `HDLC`, respectively. The diabetes status was derived from the "`DIQ_J`" dataset, with a binary indicator being assigned based on the given conditions. The next phase involved merging these individual datasets using a full join operation on the `SEQN` identifier, ensuring that each subject's records across various health domains were aligned. Additional processing included the creation of `SYSBP_UT`, which reflects untreated systolic blood pressure by setting the value to 0 if the subject is on hypertension medication, and `SYSBP_T`, which captures treated systolic blood pressure with the same logic applied inversely. Finally, the `SEQN` column, which served as the unique identifier for joining the datasets, was removed to finalize the dataset. This curated dataset now serves as a rich, multi-dimensional resource for analyzing health patterns and outcomes within the NHANES cohort.

Non-simulated transportability analysis

This report draws upon key insights from the paper *"Transporting a prediction model for use in a new target population"*. A pivotal aspect highlighted in the paper is the flexibility of model application irrespective of its precise specification.

Specifically, the authors note that their results do not hinge on the prediction model being perfectly aligned with the true conditional expectation of the outcome in the source population, denoted as $E[Y|X, S = 1]$. In other words, the model $g_{\hat{\beta}}(X)$ does not necessarily have to converge to this true conditional expectation for it to be effective in application. (Steingrimsson et al. 2022)

In light of this understanding, the current report will focus on utilizing the provided logistic regression model as a tool to examine the efficacy of both simulated and non-simulated transportability analyses. This approach is intended to assess the model's performance across different population settings, underscoring the practical implications of the model's transportability, as suggested by the referenced paper. Such an exploration aims to provide a comprehensive understanding of the model's applicability and robustness, particularly in scenarios where exact model specification may not align with the underlying population dynamics.

The below Table 1 presents a statistical comparison of various covariates between men and women from the Framingham dataset. Notable differences are observed across several biometric and health-related parameters. For instance, High-Density Lipoprotein Cholesterol (`HDLC`) levels are significantly higher in women (mean=53, SD=16) compared to men (mean=44, SD=13), with a p-value less than 0.001 indicating statistical significance. Total Cholesterol (`TOTCHOL`) follows a similar trend, with women exhibiting higher mean levels (246) than men (226), also with a p-value below 0.001. Cardiovascular Disease (`CVD`) prevalence shows a significant gender disparity, with 33% of men affected versus 17% of women, and the Body Mass Index (`BMI`) is slightly lower in women (mean=25.5, SD=4.2) than in men (mean=26.2, SD=3.5), both with p-values less than 0.001. Blood pressure medication usage (`BPMEDS`) is reported in 18% of women compared to 11% in men, indicating a higher prevalence among the female participants. When looking at smoking status (`CURSMOKE`), a higher percentage of men (39%) reported being current smokers than women (31%). The systolic blood pressure taken at the time of the survey (`SYSBP_T`) and the usual treatment (`SYSBP_UT`) showed higher averages in women than men, with both parameters showing statistical significance. Diabetes prevalence (`DIABETES`) is slightly higher in men at 8.8% compared to 6.6% in women, with a p-value of 0.037, suggesting a modest but significant gender difference. `Age` and standard systolic blood pressure (`SYSBP`) showed no significant differences between genders, with p-values of 0.13 and 0.6, respectively. These findings suggest a distinct pattern of cardiovascular and metabolic risk factors between men and women in the dataset, highlighting the importance of gender-specific analysis.

Table 1: Statistis value of covariates in framingham dataset

Covariates	Men, N = 1094	Women, N=1445	p-value
HDLC	44 (13)	53 (16)	<0.001
TOTCHOL	226 (41)	246 (46)	<0.001
CVD	360 / 1,094 (33%)	242 / 1,445 (17%)	<0.001
BMI	26.2 (3.5)	25.5 (4.2)	<0.001
BPMEDS	123 / 1,094 (11%)	259 / 1,445 (18%)	<0.001
SYSBP_T	18 (51)	28 (62)	<0.001
CURSMOKE	425 / 1,094 (39%)	445 / 1,445 (31%)	<0.001
SYSBP_UT	121 (47)	111 (56)	<0.001
DIABP	82 (11)	80 (11)	<0.001
DIABETES	96 / 1,094 (8.8%)	95 / 1,445 (6.6%)	0.037
AGE	60 (8)	61 (8)	0.13
SYSBP	139 (21)	140 (24)	0.6

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

[2] Pearson's Chi-squared test; Wilcoxon rank sum test

Based on the data presented in the preceding tables, a notable disparity is observed between male and female participants. To further investigate this, employing histograms can be an effective method. Histograms will enable us to delve into the empirical distribution of continuous variables across genders. The histograms provided offer a visual comparison of the distribution of key variables within the dataset, segmented by gender. A cursory analysis reveals distinct patterns in the distribution of age, systolic blood pressure (SYSBP), diastolic blood pressure (DIABP), high-density lipoprotein (HDL), body mass index (BMI), and total cholesterol (TOTCHOL) between male and female participants. The subsequent histograms specifically illustrate the data for male participants:

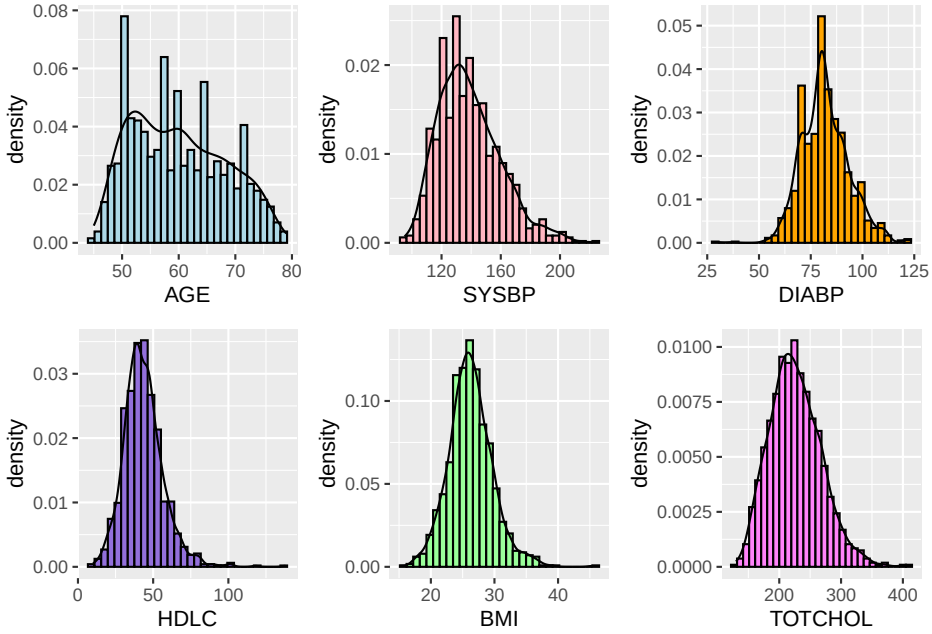

Figure 1: Male's characteristics in Framingham dataset

The series of histograms display the distribution of various health-related measurements within the Framingham dataset. Starting with age, the distribution appears approximately normal, albeit with a slight left

skew, indicating a younger population. The HDLC levels are presented with a roughly symmetrical distribution around the median, suggesting a normal variance in this cohort's cholesterol levels. In contrast, both systolic and diastolic blood pressure show a right skew, indicating that higher blood pressure levels are more variably distributed than lower levels. The Body Mass Index histogram follows a normal distribution but with a noticeable right skew, which could suggest a prevalence of higher BMI values within the population. Lastly, the total cholesterol levels exhibit a right skew as well, hinting at a population with a significant range of cholesterol levels but a tendency towards higher values. These distributions provide insights into the cardiovascular health profiles of the subjects in the dataset. The subsequent histograms specifically illustrate the data for female participants:

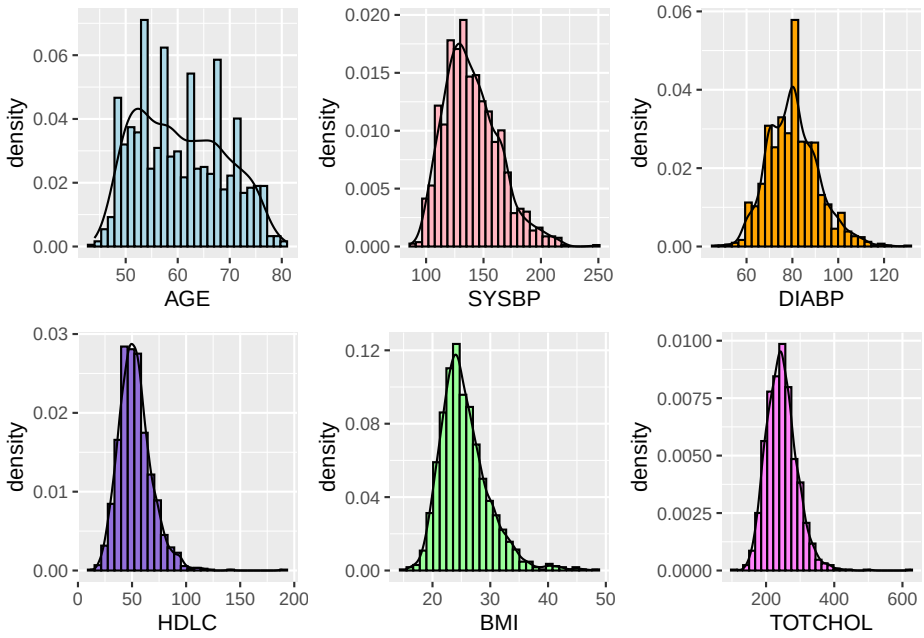

Figure 2: Female's characteristics in Framingham dataset

The above histograms offer a visual representation of key health indicators for female participants within the Framingham dataset. The age distribution is somewhat normally distributed but shows a slight leftward skew, indicating a younger median age within the sample. HDLC levels display a normal distribution, centralizing around the median value, which suggests a consistent range of cholesterol levels among the participants. SYSBP reveals a pronounced right skew, suggesting that higher blood pressure readings are more prevalent than lower ones among the females in the study. BMI also follows a normal distribution with a subtle right skew, pointing to a slightly higher occurrence of above-average BMI values. Lastly, TOTCHOL levels exhibit a distribution with a notable right skew, indicating variability with a tendency towards higher cholesterol levels in the population.

Missing values

For the source dataset (*i.e.,* framingham dataset) we just remove all the missing values for further analysis. When conducting transportability analysis, it's essential to ensure that the data accurately represent the underlying relationships and patterns we are studying. Missing values can introduce bias or distort these relationships, leading to inaccurate conclusions. By removing entries with missing values, we maintain the statistical integrity of the dataset, ensuring that all analyses are based on complete and reliable information.

Missing percentage of variables in NHANES dataset

SYSBP	DIABP	BMI	HDLC	CURSMOKE	BPMEDS	TOTCHOL	DIABETES
31.9	31.9	13.497	27.188	36.719	36.719	27.188	3.901

Based on the data presented in the preceding table, it is evident that the target dataset exhibits an excessively high percentage of missing covariate values, rendering the application of multiple imputation methods impractical. Meanwhile, our focus has shifted to utilizing non-nested results, which implies that the covariate distribution within the target dataset does not significantly impact the transportability analysis outcomes. Given this context, we have opted to eliminate all instances of missing values from the target dataset to ensure a more convenient analysis. In transportability analysis, we often compare or combine data from different sources or populations. If these datasets have varying degrees of missingness or have missing values in different variables, it can complicate the analysis. By removing all instances with missing values, we ensure a level of consistency across datasets, making it easier to conduct comparative analyses.

The below Table 3 presents the statistical values of various covariates from the NHANES dataset, comparing results between men and women. The table includes standard health and lifestyle metrics such as HDLC (High-Density Lipoprotein Cholesterol), DIABP (Diastolic Blood Pressure), SYSBP (Systolic Blood Pressure), and others like BMI (Body Mass Index) and AGE. Notably, the sample sizes are substantial, with 2105 men and 2205 women included. Across most parameters, significant differences between genders are indicated by p-values less than 0.001, suggesting that these variations are unlikely to be due to chance. Current smoking rates and diabetes prevalence, for example, are higher in men than women, as indicated by their respective p-values. However, no significant difference is noted in BMI, with a p-value of 0.5, and similarly, the use of blood pressure medication is not significantly different between genders, indicated by a p-value of 0.6.

Table 3: Statistis value of covariates in NHANES dataset

Covariates	**Men, N = 2105**	**Women, N=2205**	**p-value**
HDLC	48 (14)	58 (16)	<0.001
DIABP	73 (12)	70 (14)	<0.001
SYSBP	126 (17)	124 (20)	<0.001
SYSBP_UT	86 (58)	84 (55)	<0.001
TOTCHOL	183 (42)	191 (41)	<0.001
CURSMOKE	429 / 2,105 (20%)	316 / 2,205 (14%)	<0.001
DIABETES	370 / 2,105 (18%)	271 / 2,205 (12%)	<0.001
AGE	50 (19)	49 (19)	0.025
BMI	29 (6)	30 (8)	0.5
BPMEDS	627 / 2,105 (30%)	640 / 2,205 (29%)	0.6
SYSBP_T	40 (62)	40 (64)	>0.9
CVD	0 / 0 (NA%)	0 / 0 (NA%)	
(Missing)	2,105	2,205	

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

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

The target data, was divided into a training set and a testing set, with the training set comprising 70% of the data and the testing set the remaining 30%. Within the testing set, a total of 762 patients from the Framingham dataset were included, of which 577 patients did not experience cardiovascular disease during the follow-up period, and 185 patients did. Conversely, the training set consisted of 1,777 patients, also from the Framingham dataset, with 417 patients having developed cardiovascular disease during the follow-up period and 1,360 not having the condition occur. This division ensures a balanced representation of both outcomes – occurrence and non-occurrence of cardiovascular disease during the follow-up period – in both the training and testing datasets.

Brier score (MSE for binary outcome)

Let Y represent the outcome and X the set of covariates. We define S as the population indicator, where $S = 1$ for the source dataset and $S = 0$ for the target dataset. The function $g(X)$ denotes the prediction model for the probability $\Pr[Y = 1|X, D = 0]$. Additionally, D serves as an indicator of the dataset's division into the test set ($D = 1$) and the training set ($D = 0$). The estimated Brier risk for the target population is calculated using the following formula:

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

Here, $\hat{o}(X)$ is the estimator for the inverse-odds weighting in the test set, defined as $\frac{\Pr[S=0|X,D_{\text{test}}=1]}{\Pr[S=1|X,D_{\text{test}}=1]}$. This component adjusts the Brier risk calculation to account for the distributional differences between the source and target populations within the test set, thereby facilitating a more accurate estimation of the model's predictive performance. (Steingrimsson et al. 2022)

In our study, we constructed two separate predictive models for cardiovascular disease (CVD) risk, specifically tailored for men and women, using logistic regression. For the male population, the model was built using the train dataset which only contained male participants. The predictors included in this model are the natural logarithms of High-Density Lipoprotein Cholesterol (HDLC), Total Cholesterol (TOTCHOL), Age (AGE), untreated Systolic Blood Pressure (SYSBP_UT), and treated Systolic Blood Pressure (SYSBP_T). The model also includes binary variables for current smoking status (CURSMOKE) and diabetes status (DIABETES). The addition of 1 to the systolic blood pressure variables before taking the logarithm ensures computational stability by avoiding the logarithm of zero. Similarly, for the female population, the model was constructed using train dataset which only contained female participants. The same set of predictors was used to maintain consistency in evaluating CVD risk across genders. This approach facilitates a gender-specific assessment, acknowledging potential biological and physiological differences between men and women in CVD risk factors. Both models aim to estimate the probability of developing CVD based on these risk factors. By employing logistic regression, we quantify the odds of CVD occurrence as a function of the chosen predictors, enabling a nuanced understanding of how these factors contribute to CVD risk in different genders. This modeling approach is pivotal in identifying high-risk individuals and informing targeted preventive strategies.

For Male:

$$\text{logit}(\Pr(\text{CVD})) = \beta_0 + \beta_1 \cdot \log(\text{HDLC}) + \beta_2 \cdot \log(\text{TOTCHOL}) + \beta_3 \cdot \log(\text{AGE}) + \beta_4 \cdot \log(\text{SYSBP UT} + 1)$$
$$+\beta_5 \cdot \log(\text{SYSBP T} + 1) + \beta_6 \cdot \text{CURSMOKE} + \beta_7 \cdot \text{DIABETES}$$

For Female:

$$\text{logit}(\Pr(\text{CVD})) = \alpha_0 + \alpha_1 \cdot \log(\text{HDLC}) + \alpha_2 \cdot \log(\text{TOTCHOL}) + \alpha_3 \cdot \log(\text{AGE}) + \alpha_4 \cdot \log(\text{SYSBP UT} + 1)$$
$$+\alpha_5 \cdot \log(\text{SYSBP T} + 1) + \alpha_6 \cdot \text{CURSMOKE} + \alpha_7 \cdot \text{DIABETES}$$

A critical step involved calculating the inverse odds weights for the test dataset, which is a fundamental process in transportability analysis and other observational studies where sample selection bias might be a concern. The calculation was carried out through several stages using logistic regression and probability transformations. Initially, we fitted a logistic regression model, to the test dataset, excluding certain columns that were either outcome or collinear variable for this analysis. The model was specified to predict the binary variable S_i, which indicates whether an individual is part of the study sample or the target population. Upon fitting the model, we employed the predict function to compute the predicted probabilities of each individual being in the study sample ($S_i = 1$). This prediction was based on the covariates included in the logistic regression model. Subsequently, we transformed these probabilities into odds. The odds of an event is the probability of the event occurring divided by the probability of the event not occurring. This transformation was achieved using the formula odds $= \frac{\text{probability}}{(1-\text{probability})}$. The final and crucial step involved calculating the inverse odds weights. These weights play a vital role in balancing the distribution of the covariates between

the study sample and the target population, thereby mitigating selection bias. For individuals in the study sample ($S_i = 1$), the inverse odds weight was calculated as the reciprocal of the odds of being in the study sample. Conversely, for individuals in the target population ($S_i = 0$), the weight was calculated as the reciprocal of one minus the odds of being in the study sample. These weights are instrumental in subsequent statistical analyses, ensuring that the results are more representative of the target population.

Table 4: Estimated Brier Score in the NHANES Dataset

Brier score for women	Brier score for men
0.0509048	0.0930277

The Brier risk, commonly used to assess the accuracy of probabilistic predictions, is a measure of how well a predictive model performs. It is particularly useful in evaluating models that output probabilities for binary outcomes. The Brier score ranges from 0 to 1, with lower values indicating better prediction accuracy. A Brier score of 0 means perfect prediction, while a score of 1 indicates the worst possible prediction.(Steyerberg et al. 2010) For women, the Brier score is estimated at 0.0509048, while for men, it is slightly higher at 0.0930277. This differential suggests that the predictive model is more accurate for women than for men within this dataset. The lower score for women implies that the predictions made by the model are closer to the actual outcomes as compared to those made for men. This could be due to a variety of factors, such as differences in the variability of the outcomes being predicted, the distribution of predictive factors between genders, or potentially a model that is better tuned to the characteristics present in the female subset of the NHANES dataset.

Simulated transportability analysis

In order to simulate the `NHANES` dataset, we decided to generate all continuous variable (i.e., TOTCHOL, AGE, SYSBP, DIABP, HDLC, and BMI) follow to multivariate normal distribution. The covariance matrix (Σ) of these continuous variables from the `NHANES` dataset is meticulously computed. This matrix is crucial as it encapsulates the variance of each individual variable and the covariance amongst them, effectively preserving the inter-variable relationships observed in the original dataset.

Based on the information provided from the analysis of plot 1 and plot 2, it is evident that the variable `AGE` exhibits a uniform distribution for both male and female groups within the Framingham dataset. Specifically, the distribution of `AGE` for females ranges uniformly between 44 and 81 years, whereas for males, the uniform distribution ranges between 45 and 79 years. The variables `TOTCHOL` (total cholesterol), `HDLC` (high-density lipoprotein cholesterol), and `SYSBP` (systolic blood pressure) demonstrate a significant right-skewed distribution. A right-skewed distribution is characterized by a longer tail on the right side, indicating that a minority of higher values stretches out to the right, away from the majority of the data points. Given the right-skewed nature of these distributions, it is appropriate to consider gamma distributions to model `TOTCHOL`, `HDLC`, and `SYSBP`.

For the female cohort, systolic blood pressure is best modeled by a gamma distribution with a shape parameter of 36.260031 and a scale parameter of 3.859235 (*i.e.,* $\Gamma(36.260031, 3.859235)$). In contrast, the male cohort's systolic blood pressure follows a gamma distribution with a slightly higher shape parameter of 46.21877 and a lower scale of 3.00623 (*i.e.,* $\Gamma(46.21877, 3.00623)$), indicating a more pronounced right-skewness compared to females. When considering total cholesterol levels, the female group is characterized by a gamma distribution with a shape of 30.674790 and scale of 8.029993 (*i.e.,* $\Gamma(30.674790, 8.029993)$), whereas the male group displays a similar shape parameter at 30.372507 but with a smaller scale parameter of 7.455566 (*i.e.,* $\Gamma(30.372507, 7.455566)$). This suggests that while the concentration and variability of total cholesterol are somewhat similar between genders, the male distribution is slightly more concentrated around the mean.

For HDLC, the female gamma distribution is defined by a shape parameter of 12.051925 and a scale parameter of 4.403724 (*i.e.,* $\Gamma(12.051925, 4.403724)$), indicating a certain degree of right-skewness. The male HDLC

levels, however, are modeled by a gamma distribution with a slightly lower shape parameter of 10.936868 and a scale parameter of 3.989408 (*i.e.,* $\Gamma(10.936868, 3.989408)$), suggesting a narrower and more skewed distribution than that of females.

For this purpose, we consider each continuous variable ($X_{\text{continous},i}$) as a random variable distributed according to a multivariate normal distribution, formally denoted as $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ represents the mean vector and Σ the covariance matrix for framingham dataset. In our approach to Bernoulli variables, we generated these variables independently, rather than conditioning them on other variables. Specifically, `CURSMOKE`, `DIABETES`, `BPMEDS`, and `SEX` were each generated following a Bernoulli distribution. The probabilities for these distributions were determined based on the summary data presented in Table 3. This approach ensures that each variable is treated as a distinct entity, not influenced by the values of the others in the dataset. This disadvantage could be improved in further work.

Table 5: Statistis value of covariates in Simulated NHANES dataset

Covariates	**Men, N = 2443**	**Women, N=2557**	**p-value**
SYSBP	125 (19)	125 (19)	0.037
SYSBP_UT	90 (59)	88 (59)	0.11
CURSMOKE	424 / 2,446 (17%)	463 / 2,554 (18%)	0.5
BPMEDS	695 / 2,446 (28%)	748 / 2,554 (29%)	0.5
AGE	50 (19)	49 (19)	0.5
DIABP	71 (13)	71 (13)	0.6
SYSBP_T	35 (57)	36 (57)	0.6
BMI	29 (7)	30 (7)	0.7
TOTCHOL	186 (41)	186 (42)	0.7
DIABETES	353 / 2,446 (14%)	375 / 2,554 (15%)	0.8
HDLC	53 (15)	53 (15)	>0.9
CVD	0 / 0 (NA%)	0 / 0 (NA%)	
(Missing)	2,446	2,554	

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

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

The above Tables 3 and 4, which present summary statistics for the NHANES and simulated NHANES datasets respectively, several observations can be made. Firstly, the mean systolic blood pressure (SYSBP) shows a minor difference between men and women in both datasets, with a slightly higher mean for men in the NHANES dataset, a pattern that is also reflected in the simulated data. However, the p-value in the simulated data suggests a statistically significant difference between genders, unlike in the NHANES data. For the CURSMOKE variable, the proportion of current smokers is higher in men than women in both datasets, although the difference is less pronounced in the simulated data. Similarly, the prevalence of diabetes (DIABETES) is higher in men in both datasets, but the difference between genders is narrower in the simulated dataset. Notably, the use of blood pressure medication (BPMEDS) is consistent across both datasets, with a slightly higher usage in women, and the p-values indicate no significant difference between genders in the simulated dataset. Age distribution is nearly identical between men and women in both datasets. For other variables such as BMI, total cholesterol (TOTCHOL), and diastolic blood pressure (DIABP), the differences between the NHANES and simulated NHANES datasets are minimal, with p-values indicating non-significance in the simulated data, which suggests that the simulation process was able to recreate the NHANES distribution patterns accurately for these variables. The variables HDLC and CVD are not comparable due to missing data in the simulated dataset. Overall, the simulation appears to reflect the NHANES dataset's patterns with reasonable fidelity, though there are statistically significant differences in a few key areas, specifically in systolic blood pressure across genders.

The calculation of the Brier risk was conducted using the same methodology as previously applied in the comparison between the Framingham and NHANES datasets. The brier risk between the framinghan and

simulated dataset is 0.5337068. This score, which is lower than the NHANES comparison but still above 0, indicates a moderate level of prediction error when the model, likely developed using the Framingham data, is applied to the simulated dataset. The fact that the Brier score is lower than that for the NHANES dataset suggests that the simulated dataset may be more similar to the Framingham dataset in terms of key characteristics or risk factor distributions. This implies that the model is somewhat more transportable to the simulated dataset than to the NHANES dataset, but there's still a notable discrepancy, highlighting areas where the model could be improved or adjusted for better transportability.

Conclusion

Table 5 below offers a direct comparison of the two Brier scores. The higher brier score indicates a greater degree of inaccuracy in the model's predictions when applied to the NHANES dataset. In transportability analysis, this suggests significant challenges in applying the model developed from the Framingham dataset to the NHANES population. The differences in score could be attributed to variations in population characteristics, underlying health behaviors, socio-economic factors, or other environmental differences between the two cohorts.

Table 6: Brier Score for simulated and non-simulated transportability analysis

Between framinghan and Nhanes	Between framinghan and simulated dataset
0.6191495	0.5337068

These results demonstrate the challenges of transporting a risk prediction model from one population to another. The model shows better transportability to the simulated dataset than to the NHANES dataset, indicating that population-specific factors significantly impact model performance. The variance in Brier scores underscores the importance of understanding the underlying population dynamics and characteristics when attempting to apply a model developed in one context (Framingham) to another (NHANES or a simulated dataset). It highlights the need for caution in generalizing findings across different populations. For effective transportability, it may be necessary to recalibrate the model or incorporate additional variables that are relevant to the target population. This can help in aligning the model more closely with the specific health profiles and risk factors prevalent in the population to which the model is being transported. These insights are valuable in guiding future model development and modification efforts, ensuring that predictive models are not only accurate in their original context but also maintain their validity and reliability when applied to different populations.

Further work

In the data preprocessing phase, we opted to eliminate records with missing values from the source dataset. This approach, while streamlining the dataset, may introduce bias that could skew subsequent analyses. Future efforts should be directed towards developing a methodological framework to impute missing data. By doing so, we can preserve the integrity of the dataset and enhance the robustness of our analytical outcomes. In future research endeavors, we aim to explore a variety of simulation techniques, including bootstrap and Monte Carlo methods, to facilitate the generation of multiple iterations of diverse target datasets. This approach will enable us to robustly examine the variability and generalizability of our findings across an array of simulated scenarios.

Refereneces

Endres, Christopher. 2023. "nhanesA: NHANES Data Retrieval." https://CRAN.R-project.org/package=nhanesA.

Grembi, Jessica. 2022. "riskCommunicator: G-Computation to Estimate Interpretable Epidemiological Effects." https://CRAN.R-project.org/package=riskCommunicator.

Steingrimsson, Jon A, Constantine Gatsonis, Bing Li, and Issa J Dahabreh. 2022. "Transporting a Prediction Model for Use in a New Target Population." *American Journal of Epidemiology* 192 (2): 296–304. https://doi.org/10.1093/aje/kwac128.

Steyerberg, Ewout W., Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. 2010. "Assessing the Performance of Prediction Models." *Epidemiology* 21 (1): 128–38. https://doi.org/10.1097/ede.0b013e3181c30fb2.

Code appendix:

```
knitr::opts_chunk$set(echo = TRUE)
library("nhanesA")
library("riskCommunicator")
library("tidyverse")
library("tableone")
library("caret")
library("knitr")
library("dplyr")
library("kableExtra")
library("knitr")
library("gtsummary")
library("ggplot2")
library("gt")
library("MASS")
library("gridExtra")
library("dglm")
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)

# 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
framingham_df <- framingham_df %>%
  filter(!(CVD == 0 & TIMECVD <= 365*15)) %>%
  dplyr::select(-c(TIMECVD))
framingham_df %>%
  tbl_summary(by = SEX,
              statistic = list(all_continuous() ~ "{mean} ({sd})",
                               all_categorical() ~ "{n} / {N} ({p}%)"),
              missing_text = "(Missing)") %>%
  add_p() %>%
  #filter_p() %>%
  sort_p() %>%
  bold_labels() %>%
  modify_header(label = "**Covariates**",
                stat_1 = "**Men, N = 1094**",
                stat_2 = "**Women, N=1445**") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Statistis value of covariates in framingham dataset") %>%
```

```
  kable_styling(latex_options = "HOLD_position")
framingham_df_male = framingham_df %>% filter(SEX == 1)

p1 = ggplot(framingham_df_male, aes(x = AGE)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "lightblue") +
  geom_density()

p2 = ggplot(framingham_df_male, aes(x = SYSBP)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "lightpink") +
  geom_density()


p3 = ggplot(framingham_df_male, aes(x = DIABP)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "orange") +
  geom_density()

p4 = ggplot(framingham_df_male, aes(x = HDLC)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "mediumpurple") +
  geom_density()


p5 = ggplot(framingham_df_male, aes(x = BMI)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "palegreen1") +
  geom_density()


p6 = ggplot(framingham_df_male, aes(x = TOTCHOL)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "orchid1") +
  geom_density()

grid.arrange(p1, p2, p3, p4,p5,p6, nrow = 2, ncol = 3)
framingham_df_female = framingham_df %>% filter(SEX == 2)


pp1 = ggplot(framingham_df_female, aes(x = AGE)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "lightblue") +
  geom_density()

pp2 = ggplot(framingham_df_female, aes(x = SYSBP)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "lightpink") +
  geom_density()


pp3 = ggplot(framingham_df_female, aes(x = DIABP)) +
  geom_histogram(aes(y = ..density..),
```

```
                colour = 1, fill = "orange") +
  geom_density()


pp4 = ggplot(framingham_df_female, aes(x = HDLC)) +
  geom_histogram(aes(y = ..density..),
                colour = 1, fill = "mediumpurple") +
  geom_density()


pp5 = ggplot(framingham_df_female, aes(x = BMI)) +
  geom_histogram(aes(y = ..density..),
                colour = 1, fill = "palegreen1") +
  geom_density()

pp6 = ggplot(framingham_df_female, aes(x = TOTCHOL)) +
  geom_histogram(aes(y = ..density..),
                colour = 1, fill = "orchid1") +
  geom_density()

grid.arrange(pp1, pp2, pp3, pp4,pp5,pp6, nrow = 2, ncol = 3)
# blood pressure, demographic, bmi, smoking, and hypertension info
bpx_2017 <- nhanes("BPX_J") %>%
  dplyr::select(SEQN, BPXSY1, BPXDI1) %>%
  rename(SYSBP = BPXSY1, DIABP = BPXDI1)
demo_2017 <- nhanes("DEMO_J") %>%
  dplyr::select(SEQN, RIAGENDR, RIDAGEYR) %>%
  rename(SEX = RIAGENDR, AGE = RIDAGEYR)
bmx_2017 <- nhanes("BMX_J") %>%
  dplyr::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)) %>%
  dplyr::select(SEQN, CURSMOKE)
bpq_2017 <- nhanes("BPQ_J") %>%
  mutate(BPMEDS = case_when(
    BPQ020 == 2 ~ 0,
    BPQ040A == 2 ~ 0,
    BPQ050A == 1 ~ 1,
    TRUE ~ NA )) %>%
  dplyr::select(SEQN, BPMEDS)
tchol_2017 <- nhanes("TCHOL_J") %>%
  dplyr::select(SEQN, LBXTC) %>%
  rename(TOTCHOL = LBXTC)
hdl_2017 <- nhanes("HDL_J") %>%
  dplyr::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)) %>%
```

```r
  dplyr::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")

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)

df_2017 = df_2017 %>%
  dplyr::select(-c(SEQN))
missing_percentage = df_2017 %>%
  summarise_all(~round((sum(is.na(.)) / n()) * 100, 3))
missing_percentage = missing_percentage %>% dplyr::select(-c(SEX,AGE))

missing_percentage = missing_percentage %>% dplyr::select(-c(SYSBP_UT,SYSBP_T))
gt_tbls = gt(missing_percentage)

gt_tbls <-
  gt_tbls |>
  tab_header(
    title = "Missing percentage of variables in NHANES dataset" )

gt_tbls
df_2017 = na.omit(df_2017)

df_2017$CVD = NA

df_2017 = df_2017 %>% dplyr::select(c(13, 3, 9, 4, 1, 2, 7, 10, 8, 6, 5, 11, 12))
df_2017 %>%
  tbl_summary(by = SEX,
              statistic = list(all_continuous() ~ "{mean} ({sd})",
                               all_categorical() ~ "{n} / {N} ({p}%)"),
              missing_text = "(Missing)") %>%
  add_p() %>%
  #filter_p() %>%
  sort_p() %>%
  bold_labels() %>%
  modify_header(label = "**Covariates**",
                stat_1 = "**Men, N = 2105**",
                stat_2 = "**Women, N=2205**") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Statistis value of covariates in NHANES dataset") %>%
  kable_styling(latex_options = "HOLD_position")
# Setting seed for reproducibility
```

```r
set.seed(2550)

indices = sample(1:nrow(framingham_df), size = floor(0.7 * nrow(framingham_df)))
train_set = framingham_df[indices, ]
test_set = framingham_df[-indices, ]
train_set$D_test = 0


train_set = train_set %>% mutate(S_i = case_when(
  is.na(CVD) ~ 0,
  TRUE ~ 1
))
test_dataset = rbind(test_set, df_2017)
test_dataset$D_test = 1

test_dataset = test_dataset %>% mutate(S_i = case_when(
  is.na(CVD) ~ 0,
  TRUE ~ 1
))
# Filter to each sex
train_set$CVD = as.factor(train_set$CVD)
train_set$CURSMOKE = as.factor(train_set$CURSMOKE)
train_set$DIABETES = as.factor(train_set$DIABETES)

train_set_men <- train_set %>% filter(SEX == 1)
train_set_women <- train_set %>% 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= train_set_men, family= "binomial")


mod_women <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                   log(SYSBP_T+1)+CURSMOKE+DIABETES,
               data= train_set_women, family= "binomial")
test_dataset$SEX = as.factor(test_dataset$SEX)
test_dataset$CVD = as.factor(test_dataset$CVD)
test_dataset$CURSMOKE = as.factor(test_dataset$CURSMOKE)
test_dataset$DIABETES = as.factor(test_dataset$DIABETES)
test_dataset$BPMEDS = as.factor(test_dataset$BPMEDS)
test_dataset$S_i = as.factor(test_dataset$S_i)
test_dataset_men = test_dataset %>% filter(SEX == 1)
test_dataset_women = test_dataset %>% filter(SEX == 2)
logit_weights_men = glm(S_i ~ log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                 log(SYSBP_T+1)+CURSMOKE+DIABETES, data= test_dataset_men[-c(1,2,14)], family= "binomial

inverse_weights_train_men <- 1/(exp(predict(logit_weights_men,
                                            newdata = test_dataset_men[-c(1,2,14)])))
test_set$CURSMOKE = as.factor(test_set$CURSMOKE)
test_set$DIABETES = as.factor(test_set$DIABETES)
test_set_men = test_set %>% filter(SEX == 1)
```

```
g_X_men <- ifelse(predict(mod_men,
                          newdata = test_set_men, type = "response") > 0.5, 1, 0)

pred_weights_men <- predict(logit_weights_men, newdata=test_set_men %>% dplyr::select(-'CVD'))

inverse_weights_test_men <- 1/(exp(pred_weights_men))

brier_score_df_men <- data.frame(CVD = test_set_men$CVD, g_X = g_X_men,
                                       inverse_weights_test_men =inverse_weights_test_men)

brier_score_df_men$numerator <- (brier_score_df_men$CVD
                                        -brier_score_df_men$g_X)^2*brier_score_df_men$inverse_weights_t

brier_score_men <- sum(brier_score_df_men$numerator)/nrow(test_dataset_men)
logit_weights_women = glm(S_i ~ log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES, data= test_dataset_women[-c(1,2,14)], family= "binom

inverse_weights_train_women <- 1/(exp(predict(logit_weights_men,
                                                    newdata = test_dataset_women[-c(1,2,14)])))
test_set_women = test_set %>% filter(SEX == 2)
g_X_women <- ifelse(predict(mod_women,
                          newdata = test_set_women, type = "response") > 0.5, 1, 0)

pred_weights_women <- predict(logit_weights_women, newdata=test_set_women %>% dplyr::select(-'CVD'))

inverse_weights_test_women <- 1/(exp(pred_weights_women))

brier_score_df_women <- data.frame(CVD = test_set_women$CVD, g_X = g_X_women,
                                       inverse_weights_test_women =inverse_weights_test_women)

brier_score_df_women$numerator <- (brier_score_df_women$CVD
                                        -brier_score_df_women$g_X)^2*brier_score_df_women$inverse_weigh

brier_score_women <- sum(brier_score_df_women$numerator)/nrow(test_dataset_women)
temp_model = glm(S_i ~.,
                 data = test_dataset[,-c(1,14,13,12)], family = "binomial")
test_dataset$probability = predict(temp_model, newdata = test_dataset, type = "response")
test_dataset$odds = test_dataset$probability/(1-test_dataset$probability)
test_dataset$inverse_odds_weights = ifelse(test_dataset$S_i == 1,
                                           1 / test_dataset$odds,
                                           1 / (1 - test_dataset$odds))
test_dataset_male = test_dataset %>% filter(S_i == 1 & D_test == 1 & SEX == 1)
test_dataset_male$fitted_value = predict(mod_men, newdata = test_dataset_male, type = "response")
test_dataset_female = test_dataset %>% filter(S_i == 1 & D_test == 1 & SEX == 2)
test_dataset_female$fitted_value = predict(mod_women, newdata = test_dataset_female, type = "response")
temp_temp_dataset = rbind(test_dataset_male, test_dataset_female)
numerator_value = sum(temp_temp_dataset$inverse_odds_weights*
                        ((as.numeric(temp_temp_dataset$CVD)-temp_temp_dataset$fitted_value)^2))
denominator_value = test_dataset %>% filter(S_i == 0 & D_test == 1) %>% dim()
denominator_value = denominator_value[1]
Brier_risk = numerator_value/denominator_value
BrierScore_dataset = cbind(brier_score_women, brier_score_men)
BrierScore_dataset = as.data.frame(BrierScore_dataset)
```

```
BrierScore_dataset %>%
  mutate_all(linebreak) %>%
  kbl(caption = "Estimated Brier Score in the NHANES Dataset",
      col.names=linebreak(c("Brier score for women", "Brier score for men")),
      booktabs=T, escape=F, align = "c") %>%
kable_styling(full_width = FALSE, latex_options = c('hold_position'))
mlgamma <- function(x) {
  fit <- dglm(
    x~1,
    family=Gamma(link="log"),
    mustart=mean(x)
  )
  mu <- exp(fit$coefficients)
  shape <- exp(-fit$dispersion.fit$coefficients)
  scale <- mu/shape
  result <- c(shape, scale)
  names(result) <- c("shape", "scale")
  result
}
continous_nhanes = df_2017 %>% dplyr::select(-c(CVD, SEX, CURSMOKE, DIABETES, BPMEDS, SYSBP_UT,SYSBP_T))
cov_matrix = cov(continous_nhanes)
mean_vector = colMeans(continous_nhanes, na.rm = TRUE)
simulated_continous = mvrnorm(n = 5000, mu = mean_vector, Sigma = cov_matrix)
simulated_nhanes = as.data.frame(simulated_continous)
CURSMOKE_probability = sum(df_2017$CURSMOKE)/length(df_2017$CURSMOKE)
simulated_nhanes$CURSMOKE = rbinom(n = 5000, size = 1, prob = CURSMOKE_probability)
DIABETES_probability = sum(df_2017$DIABETES)/length(df_2017$DIABETES)
simulated_nhanes$DIABETES = rbinom(n = 5000, size = 1, prob = DIABETES_probability)
BPMEDS_probability = sum(df_2017$BPMEDS)/length(df_2017$BPMEDS)
simulated_nhanes$BPMEDS = rbinom(n = 5000, size = 1, prob = BPMEDS_probability)
simulated_nhanes$CVD = NA
SEX_probability = (sum(df_2017$SEX) - length(df_2017$BPMEDS))/length(df_2017$BPMEDS)
simulated_nhanes$SEX = rbinom(n = 5000, size = 1, prob = SEX_probability)
simulated_nhanes$SEX = simulated_nhanes$SEX+1
simulated_nhanes$SYSBP_UT <- ifelse(simulated_nhanes$BPMEDS == 0,
                                    simulated_nhanes$SYSBP, 0)
simulated_nhanes$SYSBP_T <- ifelse(simulated_nhanes$BPMEDS == 1,
                                   simulated_nhanes$SYSBP, 0)
simulated_nhanes %>%
  tbl_summary(by = SEX,
              statistic = list(all_continuous() ~ "{mean} ({sd})",
                               all_categorical() ~ "{n} / {N} ({p}%)"),
              missing_text = "(Missing)") %>%
  add_p() %>%
  #filter_p() %>%
  sort_p() %>%
  bold_labels() %>%
  modify_header(label = "**Covariates**",
                stat_1 = "**Men, N = 2443**",
                stat_2 = "**Women, N=2557**") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Statistis value of covariates in Simulated NHANES dataset") %>%
  kable_styling(latex_options = "HOLD_position")
```

```r
simulated_nhanes = simulated_nhanes %>% dplyr::select(c(10, 11, 1, 2, 3, 4, 7, 8, 9, 5, 6, 12, 13))
test_simulate_dataset = rbind(test_set, simulated_nhanes)
test_simulate_dataset$D_test = 1

test_simulate_dataset = test_simulate_dataset %>% mutate(S_i = case_when(
  is.na(CVD) ~ 0,
  TRUE ~ 1
))
test_simulate_dataset$SEX = as.factor(test_simulate_dataset$SEX)
test_simulate_dataset$CVD = as.factor(test_simulate_dataset$CVD)
test_simulate_dataset$CURSMOKE = as.factor(test_simulate_dataset$CURSMOKE)
test_simulate_dataset$DIABETES = as.factor(test_simulate_dataset$DIABETES)
test_simulate_dataset$BPMEDS = as.factor(test_simulate_dataset$BPMEDS)
test_simulate_dataset$S_i = as.factor(test_simulate_dataset$S_i)
temp_simulate_model = glm(S_i ~.,
                data = test_simulate_dataset[,-c(1,14,13,12)], family = "binomial")
test_simulate_dataset$probability = predict(temp_model, newdata = test_simulate_dataset, type = "respons
test_simulate_dataset$odds = test_simulate_dataset$probability/(1-test_simulate_dataset$probability)
test_simulate_dataset$inverse_odds_weights = ifelse(test_simulate_dataset$S_i == 1,
                                        1 / test_simulate_dataset$odds,
                                        1 / (1 - test_simulate_dataset$odds))

test_simulate_male = test_simulate_dataset %>% filter(S_i == 1 & D_test == 1 & SEX == 1)
test_simulate_male$fitted_value = predict(mod_men, newdata = test_simulate_male, type = "response")

test_simulate_female = test_simulate_dataset %>% filter(S_i == 1 & D_test == 1 & SEX == 2)
test_simulate_female$fitted_value = predict(mod_women, newdata = test_simulate_female, type = "response

temp_simulate = rbind(test_simulate_male, test_simulate_female)
temp_numerator_value = sum(temp_simulate$inverse_odds_weights*
                        ((as.numeric(temp_simulate$CVD)-temp_simulate$fitted_value)^2))


temp_denominator_value = test_simulate_dataset %>% filter(S_i == 0 & D_test == 1) %>% dim()
temp_denominator_value = temp_denominator_value[1]
simulated_brier_risk = temp_numerator_value/temp_denominator_value
BrierScore_dataset = cbind(Brier_risk, simulated_brier_risk)
BrierScore_dataset = as.data.frame(BrierScore_dataset)
BrierScore_dataset %>%
  mutate_all(linebreak) %>%
  kbl(caption = "Brier Score for simulated and non-simulated transportability analysis",
      col.names=linebreak(c("Between framinghan and Nhanes", "Between framinghan and simulated dataset"
      booktabs=T, escape=F, align = "c") %>%
kable_styling(full_width = FALSE, latex_options = c('hold_position'))
```