Assessing the Transportability of a Cardiovascular Disease Event Model Constructed from the Framingham Heart Study to the NHANES Dataset

Project 3: Simulation Studies*

Haiyue Song

2023-12-12

Abstract

Background: Assessing transportability is of importance when users of prediction models are interested in applying model-derived predictions to some different target populations. In recent years, several methods have been developed to evaluate the performance of prediction models in a target population (Paul 2023; Steingrimsson et al. 2023). In this study, we are interested in transporting the model for cardiovascular disease events constructed from Framingham Heart Study data to the more extensive National Health and Nutrition Examination Survey data.

Methods: When the target population's covariates were available, we utilized multiple imputation to address missing data and then calculated the Brier score in NHANES data with both the original Framingham model and the tailored model, straitified by sex. Additionally, when only summary data were available from the target population, simulation studies are conducted by drawing random samples using the summary statistics of the target population, and the Brier score was then calculated.

Results: The Brier score of original Framingham model averages 0.1085 for men and 0.0412 for women, and the Brier score of tailored model averages 0.1056 for men and 0.0441 for women in NHANES population, which are about 50% of the Brier score validated on Framingham data themselves. The results of Brier score from simulation studies are close to the Brier score calculated using individual-level data directly.

Conclusions: The results indicate the excellent performance of transporting the model constructed from Framingham data to the national NHANES data. Moreover, the simulation studies indicate when only summary data are available, the transportability analysis is workable and promising.

Keywords: Transportablity, Cardiovascular Disease Events, Framingham Heart Study, NHANES

1. Introduction

Users of prediction models are usually interested in applying model-derived predictions to some target populations. A healthcare system might use an existing risk prediction model to identify individuals at high risk for cardiovascular events among all patients in its care. For example, some are interested in using the risk prediction model built from the Framingham Heart Study in the extensive population. However, the data used for developing the prediction models, often referred to as the source study data, are not typically random samples from the target population. Consequently, there are differences between the target population and the population underlying the source study data (Paul 2023).

*This project is a collaboration with Dr. Jon Steingrimsson in the Department of Biostatistics, Brown University School of Public Health. The instructor is Dr. Alice Paul from Department of Biostatistics, Brown University School of Public Health.

In recent years, several methods have been developed to evaluate the performance of prediction models in a target population. Steingrimsson et al. (2023) discussed how to tailor the prediction model to account for differences in the data distribution between the source population and the target population, when the covariates for the target population is available. They also discuss how to assess the model's performance (e.g., by estimating the mean squared prediction error) in the target population.

In this study, we developed a prediction model for cardiovascular disease (CVD) events, using data from the Framingham Heart Study. Separate models were created for men and women. We assessed the model's transportability for application to a new target population, represented by the National Health and Nutrition Examination Survey (NHANES) data, employing the Brier score as a measure of transportability based on the method described by Steingrimsson et al. (2023). We also evaluated the transportability of tailored models. Our results show excellent transportability of the Framingham-based CVD event prediction models for men and women transported to the national NHANES data. Additionally, when only summary data was available for NHANES data, simulation studies are conducted by drawing random samples using the summary statistics of the target population. Some simulated data gave promising results, indicating that the transportability analysis is workable when individual-level data are inaccessible.

2. Methods

2.1. Setup and Notation

Y denotes the outcome of interest. $\mathbf{X}$ denotes the covariates. Framingham study is the source population with both covariates and outcome available as $\{(X_i, Y_i), i = 1, ..., n_{\text{source}}\}$; NHANES is the target population with only covariates available as $\{X_i, i = 1, ..., n_{\text{target}}\}$.

Let S denote the population indicator ($S = 1$ if in the source Framingham study and $S = 0$ if in the target NHANES). The composite data consisting of both source population and target population has $n = n_{\text{source}} + n_{\text{target}}$ observations. This composite data set is randomly split into a training set and a test set. Let D denote an indicator for test set ($D = 1$ if in test set and $D = 0$ if in training data).

$g_\beta(\mathbf{X}) = Pr(Y = 1|X, D = 0)$ is a parametric prediction model trained on training data. Then, $g_{\hat{\beta}}(\mathbf{X})$ denotes the fitted model with estimated $\hat{\beta}$. The MSE is denoted by ψ.

2.2. Data

2.2.1 Source Population

The source population comprises patients from the Framingham Heart Study, a long-term prospective study investigating the etiology of cardiovascular disease (CVD) among a population of free-living individuals in Framingham, Massachusetts ("Framingham Heart Study" 2023). D'Agostino et al. (2008) used a regression model to assess the risk of developing CVD events, including coronary heart disease, cerebrovascular disease, peripheral vascular disease, and heart failure. The models were developed based on sex, according to previous studies. Risk factors considered included age, serum total cholesterol (Total-C), high-density lipoprotein cholesterol (HDL-C), systolic blood pressure (Systolic BP), use of anti-hypertensive medication during the exam (BP treatment), current cigarette smoking status at the time of the exam, diabetes status, and Body Mass Index (BMI) (D'Agostino et al. 2008).

The data were obtained from the R package *riskCommunicator* (Grembi and McQuade 2022). Entries with missing data were excluded. There are 2,539 observations in total, comprising 1,094 men and 1,445 women. We created two new variables for Systolic BP, contingent on BP treatment: 'Systolic BP with treatment', which is equal to the patient's Systolic BP if they are on BP medication, otherwise, it is 0; and 'Systolic BP without treatment', which is equal to the patient's Systolic BP if they are not on BP medication, otherwise, it is 0. The dataset was filtered to include only those with a minimum of 15 years of follow-up, excluding censored data. Stratied by sex, 80% of the records were randomly selected for the training data and 20% for the test data. Of the 1094 male observations, 876 were allocated to the training data and 218 to the test data; of the 1445 female observations, 1156 were allocated to the training data and 289 to the test data.

2.2.2. Target Population

The target population comprises patients in the National Health and Nutrition Examination Survey (NHANES) study, a program designed to assess the health and nutritional status of adults and children in the United States ("NHANES - About the National Health and Nutrition Examination Survey" 2023). The data were obtained from the R package *nhanesA* (Endres 2023). For simplicity, we subsetted the NHANES data from 2017 to 2018. We obtained variables as same as the risk factors identified in the Framingham study, but our dataset does not include the outcome of interest, CVD events. The subset of NHANES 2017 to 2018 participants aged 30 to 74 years was used to align with the eligibility criteria of the Framingham study.

There are 4,060 entries in total, with 1,961 men and 2,099 women. A train-test split was performed, stratified by sex. Of the 1,961 male observations, 1,569 were included in the training data and 392 in the test data; of the 2,099 female observations, 1,680 were included in the training data and 419 in the test data.

For the missing data, 645 (15.89%) of Systolic BPs, 238 (5.86%) of BMI measurements, 427 (10.52%) of HDL-C, 248 (6.11%) of BP treatment records, 427 (10.52%) of Total-C, and 1 (0.02%) of diabetes status records are missing. Given these acceptable percentages of missing data and after exploratory data analysis, we assume the data is Missing At Random (MAR) and used multiple imputation to fill in the missingness before assessing transportability (Rubin 1976, 2018). Patient ID was excluded from the imputation process. We performed multiple imputation on the training data using the *mice* function (Buuren and Groothuis-Oudshoorn 2011), generating five complete datasets. This imputation method was then consistently applied to the test data, resulting in five complete test datasets.

2.2.3. Summary Statistics

Table 1 presents the summary statistics for the pre-processed data without imputation and transformation. In the Framingham source population, 360 (33%) of men experienced cardiovascular disease (CVD) events, compared to 242 (17%) of women. Women showed higher standard deviations and wider ranges than men in Total Cholesterol (Total-C), systolic blood pressure (systolic BP), High-Density Lipoprotein Cholesterol (HDL-C), and Body Mass Index (BMI). Additionally, a greater proportion of men were smokers and had diabetes, while more women were on anti-hypertension treatment. These differences between men and women highlight the need for models constructed separately for each gender. In the target population NHANES, unlike the Framingham data, there were no obvious or consistent differences in standard deviations and ranges for Total-C, systolic BP, HDL-C, and BMI between women and men. Similar to the Framingham data, more males were smokers and had diabetes in the NHANES population; however, a higher proportion of males were on anti-hypertension treatment. Missing data exites in NHANES study, and multiple imputation

Table 1: Summary Statistics

	Source		Target	
Variable	Male, N = 1,094	Female, N = 1,445	Male, N = 1,961	Female, N = 2,099
CVD	360 / 1,094 (33%)	242 / 1,445 (17%)		
Mean (SD)			NA (NA)	NA (NA)
(Missing)			1,961	2,099
Total-C				
Mean (SD)	226.44 (41.49)	246.32 (45.51)	188.40 (41.68)	195.64 (40.13)
(Missing)			218	209
AGE				
Mean (SD)	60.01 (8.18)	60.55 (8.40)	52.92 (12.71)	51.91 (12.56)
Systolic BP				
Mean (SD)	138.94 (20.89)	139.94 (23.71)	128.29 (17.67)	125.78 (20.24)
(Missing)			292	353
Current Smoking Status	425 / 1,094 (39%)	445 / 1,445 (31%)	474 / 1,961 (24%)	337 / 2,099 (16%)
Diabetes Status	96 / 1,094 (8.8%)	95 / 1,445 (6.6%)	358 / 1,961 (18%)	317 / 2,098 (15%)
(Missing)			0	1
BP Treatment	123 / 1,094 (11%)	259 / 1,445 (18%)	604 / 1,829 (33%)	629 / 1,983 (32%)
(Missing)			132	116
HDL-C				
Mean (SD)	43.63 (13.37)	53.07 (15.67)	47.83 (14.07)	57.60 (15.84)
(Missing)			218	209
BMI				
Mean (SD)	26.25 (3.47)	25.55 (4.22)	29.90 (6.56)	30.82 (8.19)
(Missing)			125	113

[1] n / N (%)

was used to fill in the missingness. Table 1 also reveals significant disparities between the Framingham and NHANES data, indicating the importance of assessing the transportability of the prediction model from Framingham study to new target national population NHANES.

2.3. Transportability Analysis

When the covariates of the target population is available, we can transport a prediction model from source population for use in a new target population and assess the Brier score of the model in the target population. Steingrimsson et al. (2023) discussed how to tailor the prediction model to account for differences in the data distribution between the source population and the target population. They also discuss how to assess the model's performance (e.g., by estimating the mean squared prediction error) in the target population.

2.3.1. Composite Data

We composed the training data from both the source population Framingham data and the target population NHANES data. And the test data are also combined. As we set, S is the source indicator and D is the test indicator.

Table 2: Summary Statistics of Log-transformed Continuous Variables

	Source		Target	
Variable	Male, N = 1,094	Female, N = 1,445	Male, N = 1,961	Female, N = 2,099
Log-transformed Total-C				
Mean (SD)	5.41 (0.18)	5.49 (0.18)	5.21 (0.22)	5.26 (0.20)
(Missing)			218	209
Log-transformed Age				
Mean (SD)	4.09 (0.14)	4.09 (0.14)	3.94 (0.26)	3.92 (0.26)
Log-transformed Systolic BP				
Mean (SD)	4.92 (0.15)	4.93 (0.17)	4.85 (0.13)	4.82 (0.15)
(Missing)			292	353
Log-transformed HDL-C				
Mean (SD)	3.73 (0.31)	3.93 (0.29)	3.83 (0.27)	4.02 (0.27)
(Missing)			218	209

2.3.2. Model

According to D'Agostino et al. (2008), the continuous risk factors, including age, Total-C, HDL-C, Systolic BP are log-transformed involving in the regression model, and the discrete variables including BP treatment, current cigarette smoking status at the time of the exam, and diabetes status, involve in the model directly. Logistic regression is used to fit the outcome of interest, the CVD events. We will first transport this model fitted on training data from Framingham study directly to all NHANES data.

Table 2 shows the summary statistics of log-tranformed continuous variables involving in the model, stratified by sex and population.After log-transformation, the mean and standard deviation are closer between Framingham and NHANES. Those log-transformed means in target NHANES are further used to generate data in simulation studies.

2.3.3. Model Tailoring

We also considered to tailor the model before transporting to NHANES. We used the approach for tailoring a potentially misspecified prediction model for use in the target population from Steingrimsson et al. (2023). This approach includes three steps. First, estimate the propensity score of S (i.e., $Pr(S = 1|X, D = 0)$) using training data. Second, use the estimated propensity to calculate the inverse-odds weights (i.e., $\frac{Pr(S=0|X,D=0)}{Pr(S=1|X,D=0)}$) for observations in the training set from the source population. Third, use the weights from the second step to estimate the prediction model $g_{\hat{\beta}}(\mathbf{X})$ using all observations in the training set from the source population.

2.3.4. Assessing Model Performance in the Target Population

According to Steingrimsson et al. (2023), under certain conditions and assumptions, the MSE (i.e., the Brier Score for binary outcome) estimator for the target population is

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

where $\hat{o}(X_i)$ is an estimator for the inverse-odds weights in the test set, $\frac{Pr(S=0|X,D=1)}{Pr(S=1|X,D=1)}$.

2.4. Simulation Studies

2.4.1. Assumptions

We assume that only the summary data of the target population NHANES are available and the individual level data is not available. Therefore, in order to use the methods introduced in Section 2.3, we need to simulate individual level data to mimic that summary statistics. Then, transportability analysis is conducted using this simulated dataset.

We assume both the marginal and the joint distributions of covariates from target population NHANES study are within the location-scale family of the marginal and the joint distributions of covariates from source population Framingham study; we assume the correlations between covariates from target population are as same as the correlations between covariates from source population. That is, though exploratory analysis of distributions for Framgingham data, with the summary statistics for NHANES data, we can generate the individual-level data of NHANES jointly or marginally.

2.4.2 Marginal Distribution

Figure 1 shows marginal distributions for log-transformed continuous data in Framingham grouped by sex. The red line denotes the density of the data, while the blue dashed line denotes the probability density function of normal distribution with mean and standard deviation estimated from the data. The continuous variables are approximately marginally normally distributed after log-transformation. Except for the log-transformed age, the other variables are very close to a normal distribution, making random sampling from a normal distribution promising. However, the distribution of log-transformed age shows a slightly negative kurtosis, which may result in less accurate sampling. We will consider to varying data-generating mechanisms based on these results in Section 2.4.3.

2.4.3. ADEMP Framework

We use ADEMP framework (Morris et al. 2019) to develop simulation studies.

- **Aims**: The aim of this simulation study is to investigate the transportability of a prediction model for use in a new target population by Brier score, when only summary data is available. Specifically, we assess the transportability of the CVD event model built on Framingham data for use in the NHANES national population when only summary data of NHANES is available and covariates are generated based on available summary statistics.
- **Data-generating mechanisms**: We designed three data-generating mechanisms.
    1. **From a multivariate normal distribution when summary of NHANES after log-transformation to continuous variables is available (DGM1)**: after exploratory analysis, most of continuous variables approximately and marginally follow the lognormal distribution; the analysis results of distributions are shown in Section 3.2. The categorical variables are also correlated with the continuous variables; for example, BP medication treatment is obviously correlated with Systolic BP and age, and current smoking status is correlated with age. Therefore, to preserve correlations between continuous variables and categorical variables or between two categorical variables, we decided our first data-generating mechanism: drawing random samples

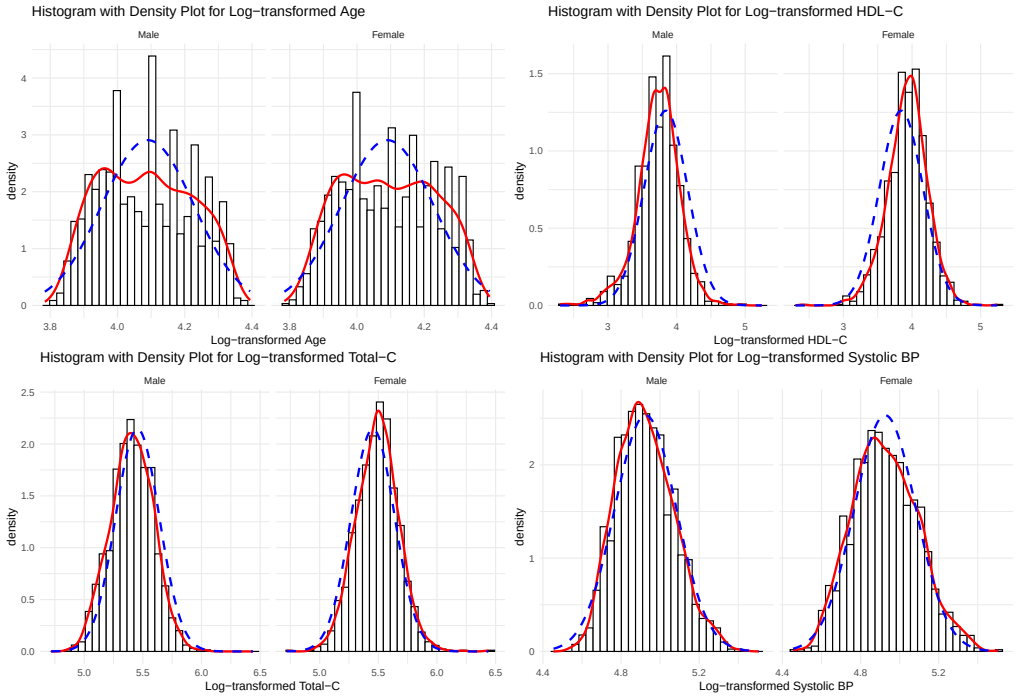

Figure 1: Marginal Distributions for Continuous Data in Framingham

from a joint multivariate normal distribution for all continuous and categorical variables, stratified by sex. Subsequently, transformations based on quantiles are conducted to convert continuous values into categorical values for categorical data.

Since we have a total of seven variables, seven-dimensional random samples are drawn from a joint multivariate normal distribution, denoted as $N_7(\hat{\mu}, \hat{\Sigma})$, where $\hat{\mu}$ is obtained from summary data for NHANES after log-transformation to continuous variables, and $\hat{\Sigma}$ is the estimated covariance matrix from summary data of NHANES after log-transformation to continuous variables. Based on the proportion of categories for categorical variables, we determine the thresholds using quantiles to categorize the data.

2. **From a multivariate normal distribution when only summary of raw data are available (DGM2):** In practice, we may only have access to the summary of raw data, instead of summary of data after log-transformation to continuous variables. In this case, we need to calculate $\hat{\mu}$ and $\hat{\sigma}$. Similarly like DGM1, we assume the continuous variables follow lognormal distribution. Let M_i and V_i denote the marginal mean and standard deviation from the summary of raw data. Referred to "Log-normal distribution" (2023), we have the equation for mean μ_i and standard deviation σ_i marginally after log-transformation to continuous variables.

$$\mu_i = \log(\frac{M_i^2}{\sqrt{V_i + M_i^2}}),$$

$$\sigma_i = \sqrt{\log(\frac{V_i}{M_i^2} + 1)},$$

where i is the subscript denoting different continuous variables. Then, the covariance matrix can be calculated as

$$\Sigma_{i,j} = \rho_{i,j}\sigma_i\sigma_j,$$

where $i, j = 1, \ldots, 7$. Finally, the other steps are as same as the DGM1.

3. **From marginal distributions with Cholesky transformation (DGM3):** In DGM1 and DGM2, we draw data from multivariate normal distributions. Except for the log-transformed age, the other variables are very close to a normal distribution, making random sampling from a normal distribution promising. However, the distribution of log-transformed age shows a slightly negative kurtosis, which may result in less accurate sampling. We then considered to analyze the marginal distribution of each continuous variable and generate the sample marginally. We finally use Cholesky transformation to make the simulated data correlated as Framingham.

   After exploratory analysis by skewness-kurtosis plots created by *descdist* function from *fitdistrplus* package (Delignette-Muller and Dutang 2015), we determined the marginal distribution of each variable. Based on Framingham data, Total-C and HDL-C follow lognormal distributions, and Systolic BP follows a gamma distribution. The parameters of the distributions can be estimated via first and second moments from summary data of NHANES. We draw random samples marginally for these continuous variables from these distributions and for categorical variables from standard normal distribution. For age, we also generate it using standard normal distribution firstly. $\hat{\Sigma}$ is the covariance matrix in DGM1. Let $\Sigma = L \cdot L^T$, where L is the lower triangular square root matrix after Cholesky decomposition. Let s denote a random sample column vector. Then, $s' = L \cdot s$ gives the correlated data after Cholesky transformation. And the age is finally map to UNIF$(30, 74)$.

Data are simulated using R 4.1.2, with the input seed "2550". The data-generating mechanisms simulate $n_{obs} = 2000$ for each sex in each repitition.

- **Estimands**: The estimand for this simulation studies is the Brier score in target population (Steingrimsson et al. 2023) introduced in Section 2.3.4:

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

  where $\hat{o}(X_i)$ is an estimator for the inverse-odds weights in the test set, $\frac{Pr(S=0|X,D=1)}{Pr(S=1|X,D=1)}$.

- **Methods**: We firstly subset the simulated data by filtering the age within 30 to 74 years old according to the eligibility criteria of Framingham, similar to Section 2.2.2. With Framingham and simulated NHANES data, we conducted transportability analysis as Section 2.3. The seed for train-test split for simulated data for transporting tailored model is "2550".

- **Performance measures**: To assess the transportability when only summary data of target population is available, we compared the estimated Brier score from simulated data and the Brier score calculated from NHANES individual-level data directly. We will assess the bias, empirical standard errors, and MSE for $\hat{\psi}_{\hat{\beta}}$. The simulation consists of 1000 repetitions.

3. Results

3.1. Transportability Analysis Using Individual-level NHANES Data

The Framingham model is fitted on Framingham's training data. The Brier score on Framingham's test data is 0.2134 for men and 0.1218 for women. The resulted Brier score in the target population NHANES are shown in Table 3, under both Framingham models and tailored models.

Table 3: Brier Scores in the Target Population by Sex

	Framingham Model		Tailored Model	
Imputation	Men	Women	Men	Women
1	0.1074	0.0408	0.1044	0.0444
2	0.1092	0.0421	0.1054	0.0463
3	0.1089	0.0407	0.1042	0.0440
4	0.1089	0.0405	0.1081	0.0419
5	0.1079	0.0420	0.1061	0.0437
Mean	0.1085	0.0412	0.1056	0.0441
SD	0.0008	0.0008	0.0016	0.0016

First, the transportability of the prediction model for women is potentially better than that for men. Second, we noticed that the Brier score for both women and men are smaller than those from Framingham before transportation, showing the models are robust and proper to use in NHANES population. This also potentially indicates that there might be serious imbalance of patients with or without CVD events in NHANES population. Third, for men, the Brier score for NHANES shows better performance under the tailored model than under the Framingham model, indicating the efficiency of model tailoring. Finally, the standard deviation of the Brier score estimates are small, showing the consistency of great transportability across multiple imputation data sets.

3.2. Simulation Studies

The summary statistics of Brier score in simulated NHANES data are show in Table 4.

Table 4: Summary for Brier Score in Simulated Target Population

vars	n	mean	sd	median	min	max	range
Framingham Model, DGM1							
Men	1000	0.097	0.004	0.097	0.084	0.111	0.027
Women	1000	0.032	0.002	0.032	0.026	0.039	0.013
Tailored Model, DGM1							
Men	1000	0.096	0.008	0.095	0.071	0.128	0.057
Women	1000	0.034	0.004	0.034	0.021	0.049	0.028
Framingham Model, DGM2							

Table 4: Summary for Brier Score in Simulated Target Population
(continued)

vars	n	mean	sd	median	min	max	range
Men	1000	0.106	0.006	0.106	0.092	0.131	0.039
Women	1000	0.035	0.002	0.035	0.028	0.042	0.014
Tailored Model, DGM2							
Men	1000	0.109	0.010	0.108	0.083	0.153	0.070
Women	1000	0.037	0.004	0.037	0.025	0.053	0.028
Framingham Model, DGM3							
Men	1000	0.032	0.008	0.031	0.015	0.081	0.066
Women	1000	0.001	0.001	0.001	0.000	0.006	0.006
Tailored Model, DGM3							
Men	1000	0.009	0.003	0.009	0.002	0.028	0.026
Women	753	0.003	0.002	0.002	0.000	0.017	0.017

Note:
The data in table are rounded to keep three decimal places. The result is based on simulated NHANES data.

We can see that Brier score from DGM1 and DGM2 based on multivariate normal distributions are averagely close to the true values of Brier score when individual-level data are available. However, DGM3 based on marginal distribution sampling and transformed to correlated data by square root of covariance matrix gives much less Brier score.

Then, compared to the Brier scores calculated based on individual-level NHANES data, we displayed the estimates of performance for the Brier scores for the target NHANES population grouped by sex in Table 5.

Table 5: Estimates of performance for Brier Score in Simulated Target Population (Monte Carlo SEs in parentheses)

	Framingham Model		Tailored Model	
DGM	Men	Women	Men	Women
Bias				
DGM 1	-0.01128 (<0.00001)	-0.00905 (<0.00001)	-0.00998 (<0.00001)	-0.01031 (<0.00001)
DGM 2	-0.00255 (<0.00001)	-0.00641 (<0.00001)	0.0033 (0.00032)	-0.00711 (<0.00001)
DGM 3	-0.0767 (<0.00001)	-0.04019 (<0.00001)	-0.09642 (<0.00001)	-0.04137 (<0.00001)
MSE				
DGM 1	0.00014 (<0.00001)	0.00009 (<0.00001)	0.00017 (0.00001)	0.00012 (<0.00001)
DGM 2	0.00004 (<0.00001)	0.00005 (<0.00001)	0.00012 (0.00001)	0.00007 (<0.00001)
DGM 3	0.00595 (0.00004)	0.00162 (<0.00001)	0.00931 (0.00002)	0.00172 (0.00001)
Empirical SE				
DGM 1	0.00413 (0.00009)	0.00183 (0.00004)	0.00826 (0.00018)	0.00377 (0.00008)
DGM 2	0.00554 (0.00012)	0.00213 (0.00005)	0.01027 (0.00023)	0.00434 (0.0001)

Table 5: Estimates of performance for Brier Score in Simulated Target Population (Monte Carlo SEs in parentheses) *(continued)*

	Framingham Model		Tailored Model	
DGM	Men	Women	Men	Women
DGM 3	0.00822 (0.00018)	0.0008 (0.00002)	0.00301 (0.00007)	0.00232 (0.00006)

Note:
The data in table are rounded to keep five decimal places. The result is based on simulated NHANES data.

Table 5 shows the bias, MSE and empirical SE for the simulated Brier scores from DGM1 and DGM2 are very small, which indicates that simulated data from summary statistics are workable and feasible to assess the transportability of the prediction model for use in a new target population NHANES when the individual-level data is unavailable. The great performance from DGM2 also indicates when only the summary of raw data is available, we are able to mimic data and conduct the transportablity analysis.

4. Discussion

In our study, we constructed a prediction model for cardiovascular disease (CVD) events using the Framingham data, which includes both covariates and outcomes. Our target population is the NHANES, representing the entire United States, but with only individual-level covariates available or, in simulation studies, just the summary data of covariates. We aim to generalize the model constructed on Framingham data to the more extensive NHANES dataset, necessitating an evaluation of the model's transportability.

When individual-level covariates for the target population are available, the Brier score of original Framingham model averages 0.1085 for men and 0.0412 for women, and the Brier score of tailored model averages 0.1056 for men and 0.0441 for women. First, the model's transportability is potentially better for women than for men. Second, the Brier score validated on Framingham data is 0.2134 for men and 0.1218 for women. A decrease of about 50% in the Brier score indicates the effective transportation of the Framingham model to the NHANES population. This could be due to the NHANES data's serious outcome imbalance, simplifying predictions, or the small inverse-odds weights. Thirdly, Table 3 shows that for men, the tailored model consistently exhibits a lower Brier score than the original model. This supports the effectiveness of model tailoring, as proposed by Steingrimsson et al. (2023). Additionally, the Brier scores are very close across five imputed datasets, demonstrating the robustness of the transportability analysis in the presence of missingness addressed by multiple imputation.

In cases where only summary data for the target population is available, we proceed by making assumptions and simulating data based on fixed summary statistics. Transportability analysis is then conducted on simulated data from three designed data-generating mechanisms: DGM1, DGM2, and DGM3. When the transformed NHANES data's covariance matrix is available, DGM1 generates data using a multivariate normal distribution, based on exploratory data analysis. In contrast, DGM2, when lacking the covariance matrix, leverages the Framingham correlation matrix and estimates the marginal standard deviation by the first and second moments to estimate the transformed NHANES covariance matrix, subsequently drawing random samples from a multivariate normal distribution. DGM1 and DGM2 gave averagely close Brier score with small bias, MSE and empirical SE compared to the true values when individual-level data are available.

This suggests under proper data-generating mechanisms, the transportability analysis is workable and efficient when only summary data are available. However, DGM3, which draws random samples from various marginal distributions and transforms the data using the square root of the covariance matrix, results in much smaller Brier scores and overestimates transportability. When only the summary data is available in practice, we shoule be careful of data-generating mechanisms. Overall, under proper data-generating mechanisms, the transportability analysis is promising when only summary data are available.

However, our study has several limitations. Firstly, we used multiple imputation assuming the data is MAR, which means the missing values can be predicted from the observed data. This assumption is untestable, and may be not hold if the data is missing not at random (MNAR), potentially affecting our estimates. Furthermore, the multiple imputation was conducted on the entire training dataset without sex stratification, which might have influenced our results. Secondly, in the generation of simulation study data, to preserve correlation, I drew random samples directly from a multivariate normal distribution for both continuous and categorical data. I then categorized the data using thresholds from the quantiles of summary statistics. However, the underlying distribution of age may not adhere to a multivariate normal distribution, and converting continuous variables into categorical ones could induce additional statistical associations, potentially impacting our estimates. Copula functions may be further considered to generate correlated data. Thirdly, when individual-level data from the target population was available, we used multiple imputation to address missingness, and the Brier score calculations were based on this imputed data. In contrast, in the simulation studies, we only used summary statistics, most likely derived from complete data. This approach makes the Brier scores only partially comparable. Additionally, the small Brier scores observed in the NHANES data for both women and men might indicate a significant imbalance in the number of patients with or without CVD events, thereby simplifying the prediction task. Moreover, in this study, only the Brier score is assessed for the transportablity; further study can involve more measures (e.g., area under the ROC curve) to enhance the reliabilility of transportablity.

5. Conclusions

In this project, we developed a prediction model for cardiovascular disease events based on Framingham data according to previous studies, creating separate models for men and women. We then assessed the model's transportability for application to a new target population, represented by the national NHANES data, using the Brier score in target population. When the target population's covariates were available, we utilized multiple imputation to address missing data and then calculated the Brier score in NHANES data with both the original Framingham model and the tailored model, straitified by sex. The Brier score of original Framingham model averages 0.1085 for men and 0.0412 for women, and the Brier score of tailored model averages 0.1056 for men and 0.0441 for women in NHANES population, indicating excellent transportability to the NHANES data. Moreover, the simulation studies are conducted to investigate whether the transportability analysis is workable when only the summary data are accessible. Our simulation studies, which drew random samples using the summary statistics of the target population, yielded Brier scores close to those calculated directly from the NHANES data under proper data-generating mechanisms. This similarity validates the efficiency of transportability analysis when only summary data of target population are available.

References

Buuren, S. van, and Groothuis-Oudshoorn, K. (2011), "Mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45, 1–67. https://doi.org/10.18637/jss.v045.i03.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. (2008), "General Cardiovascular Risk Profile for Use in Primary Care," *Circulation*, 117, 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579.

Delignette-Muller, M. L., and Dutang, C. (2015), "Fitdistrplus: An R Package for Fitting Distributions," *Journal of Statistical Software*, 64, 1–34. https://doi.org/10.18637/jss.v064.i04.

Endres, C. (2023), "nhanesA."

"Framingham Heart Study" (2023), *Wikipedia*.

Grembi, J. A., and McQuade, E. T. R. (2022), "Introducing riskCommunicator: An R package to obtain interpretable effect estimates for public health," *PLOS ONE*, 17, e0265368. https://doi.org/10.1371/journal.pone.0265368.

"Log-normal distribution" (2023), *Wikipedia*.

Morris, T. P., White, I. R., and Crowther, M. J. (2019), "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, 38, 2074–2102. https://doi.org/10.1002/sim.8086.

"NHANES - About the National Health and Nutrition Examination Survey" (2023),.

Paul, A. (2023), "Project 3: Simulation Studies."

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

Rubin, D. B. (2018), "Multiple imputation," in *Flexible Imputation of Missing Data, Second Edition*, Chapman; Hall/CRC, pp. 29–62.

Steingrimsson, J. A., Gatsonis, C., Li, B., and Dahabreh, I. J. (2023), "Transporting a Prediction Model for Use in a New Target Population," *American Journal of Epidemiology*, 192, 296–304. https://doi.org/10.1093/aje/kwac128.

Code Appendix

All code is compatible with R version 4.1.2, and can be found in the *Project 3 Simulation Studies* folder at Github.