Evaluating the Transportability of a Cardiovascular Risk Model Across Different Populations

Jialin Liu

2023-12-15

Abstract

Background: Driven by applying prediction models in specific target populations, this project aimed to assess generalizability of cardiovascular disease (CVD) risk model from the Framingham Heart Study to other target populations that only covariates are available, and conducted a simulation study by generating three different simulation cases when individual-level datasets are not available and evaluated performance based on Brier scores.

Methods: A simulation approach was used to assess the performance of a cardiovascular risk model in various simulated populations, where individual-level data was not available. The evaluation involved calculating the mean and standard deviation of estimates of Brier scores across different data generation scenarios, including varying simulation sizes. These metrics were then compared with the Brier scores obtained from the original target population, specifically the NHANES-2017 dataset, to understand the model's transportability and accuracy.

Results: The first simulation case achieved the lowest average Brier score of 0.1813 for males and 0.0326 for females, outperforming the NHANES data with scores of 0.1885 for males and 0.0408 for females.

Conclusions: The minimal relative biases in our study suggest that the simulated data effectively replicates the covariate distributions found in the NHANES sample, enabling reliable transportability analysis in the target population when individual-level data is unavailable. Furthermore, the comparative analysis across three simulation scenarios underscores the importance of including covariate associations when simulating target populations, highlighting the critical role these associations play in enhancing the accuracy of simulation studies.

Introduction

Users of prediction models want to apply the models in a specific target population. Prediction models are often developed from samples in source populations, however, models cannot be directly applied to the target population since datasets are typically not random sample from the target population, even distributions of observed variables are totally different between source and target populations(Steingrimsson et al. 2022). Consequently, models built using the data from source population are not applicable to the target population so that model performance evaluation in the source population cannot perfectly reflect performance in the target population unless using tailored prediction models as an attractive alternative to evaluate performance in the target population to achieve transportability tasks(Steingrimsson et al. 2022). In many cases, both covariates and outcome are available in source populations, whereas only covariates are available in target populations without prior information about outcomes. Under the lack of outcomes in target populations, we tailor prediction models given outcomes information from the source population and assess performance of models for datasets with covariates only based on estimated Brier risk scores.

The Framingham Heart Study

It is widely accepted that age, sex, high blood pressure, smoking, dyslipidemia, and diabetes are the major risk factors for developing cardiovascular disease (CVD). The Framingham Heart Study was a landmark long term prospective study of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts, and identified effects of risk factors(D'Agostino et al. 2008). Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. The Framingham data has been used to create models for predicting cardiovascular risk given risk factors and markers of disease, such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, diagnoses of diabetes, and medication use(D'Agostino et al. 2008). From published scholar's work, the sex-specific multivariable risk factor algorithm was created to assess and predict CVD risk. The study sample consisted of attendees of the baseline examinations free of prevalent CVD who were 30 to 74 years of age with non-missing data on covariates(D'Agostino et al. 2008). Following this work's reference, after exclusions, 2438 participants (mean age, 59 years; 1380 women) remained eligible.

In Table 1, the risk factor characteristics of men and women in our sample at the baseline examinations are significantly different at the type-I error of 0.05. In our middle-aged sample, mean levels of systolic blood pressure and the prevalences of diabetes were similar in men and women. The prevalences of cigarette smoking and use of Anti-hypertensive medication were substantially higher in women. Then we created two new variables `SYSBP_UT` and `SYSBP_T` to get systolic blood pressure based on whether participants took medication or not. As we're not interested in measurement of hazard rates, we would like to remove censored data by examining risk within 15 years. Aiming to mimic models presented in the published works, we splitted the sample data by sex and fitted the sex-specific models with respect to log transforms for all continuous variables and selected categorical variables `CURSMOKE` and `DIABETES` to predict the probability of cardiovascular disease taking place as follows:

$$\begin{aligned}\log(\frac{p}{1-p}) = {} & \beta_0 + \beta_1 * \log(HDLC) + \beta_2 * \log(TOTCHOL) \\ & + \beta_3 * \log(AGE) + \beta_4 * \log(SYSBP\_UT + 1) \\ & + \beta_5 * \log(SYSBP\_T + 1) + \beta_6 * CURSMOKE + \beta_7 * DIABETES\end{aligned}$$

where `HDLC` means cholesterol, `TOTCHOL`means serum total cholesterol, and `SYSBP` represents systolic blood pressure.

Table 1: Characteristics of Risk Factors Stratified by SEX in the Framingham Data

	Men(1)	Women(2)	P-values
n	1058	1380	
CVD (mean (SD))	0.31 (0.46)	0.15 (0.36)	<0.001
TIMECVD (mean (SD))	7300.92 (2368.86)	8016.15 (1780.62)	<0.001
SEX = 2 (%)	0 (0.0)	1380 (100.0)	<0.001
TOTCHOL (mean (SD))	226.66 (41.56)	246.16 (45.96)	<0.001
AGE (mean (SD))	59.29 (7.60)	59.63 (7.65)	0.283
SYSBP (mean (SD))	138.48 (20.85)	139.03 (23.64)	0.548
CURSMOKE = 1 (%)	422 (39.9)	445 (32.2)	<0.001
DIABETES = 1 (%)	92 (8.7)	90 (6.5)	0.052
BPMEDS = 1 (%)	114 (10.8)	241 (17.5)	<0.001
HDLC (mean (SD))	43.61 (13.49)	53.17 (15.67)	<0.001
BMI (mean (SD))	26.28 (3.47)	25.55 (4.25)	<0.001

The National Health and Nutrition Examination Survey (NHANES)

Considering the Framingham data as the source population, we use the NHANES data from 2017-2018 (National Center for Health Statistics 2020) with the same covariates as the target population. We select variables including systolic blood pressure BPXSY1, gender RIAGENDR, age RIDAGEYR, Body Mass Index BMXBMI, cigarette smoking SMQ040/SMQ020, serum total cholesterol LBXTC, HDL cholesterol LBXHDD, diabetes DIQ010, and use of Anti-hypertensive medication BPQ020/BPQ040A/BPQ050A. The we followed the same step proceeded in the source population to add two variables about systolic blood pressure with medication treatment or not, and to filter ages ranging from 30 to 74. After exclusions, 3189 participants (mean age, 52 years; 1632 women) remained eligible. In Table 2, mean levels of serum total cholesterol and HDL cholesterol in the target population at the baseline examinations were significantly higher in women. In our middle-aged sample, the prevalences of diabetes and cigarettes smoking, as well as mean levels of systolic blood pressure, were substantially higher in men.

Table 2: Characteristics of Risk Factors Stratified by SEX in the NHANES Data

	Men(1)	Women(2)	P-values
n	1557	1632	
SYSBP (mean (SD))	128.08 (17.12)	125.46 (19.88)	<0.001
SEX = 2 (%)	0 (0.0)	1632 (100.0)	<0.001
AGE (mean (SD))	52.96 (12.65)	51.88 (12.54)	0.015
BMI (mean (SD))	29.89 (6.27)	30.72 (8.20)	0.001
HDLC (mean (SD))	47.65 (13.95)	57.89 (15.96)	<0.001
CURSMOKE = 1 (%)	372 (23.9)	269 (16.5)	<0.001
BPMEDS = 1 (%)	484 (33.1)	480 (31.1)	0.262
TOTCHOL (mean (SD))	188.98 (41.91)	196.18 (40.34)	<0.001
DIABETES = 1 (%)	298 (19.1)	241 (14.8)	0.001
SYSBP_UT (mean (SD))	83.32 (59.99)	82.51 (57.33)	0.705
SYSBP_T (mean (SD))	44.13 (63.50)	42.06 (63.45)	0.373

We exclude some participants with missingness in cholesterol-related variables, BMI, and blood pressures. After omitting those missing values, we have 6% of missingness in BPMEDS and only 1 record missing in DIABETES. Then we applied multiple imputation technique to infer those missing values with 5 imputation datasets. This method is trying to handle with each missing entry by estimating multiple reliable values such as regression models, running analysis across those completed dataset, aggregating all previous analyses results and analyzing how far they spread out in terms of standard deviations and confidence intervals.

Transportability Analysis

We assume that outcome and covariate information is obtained from a simple random sample from the source population (the Framingham data, $S = 1$). Furthermore, covariate information is obtained from a simple random sample from the target population (the NHANES data in 2017, $S = 0$), and no outcome information is collected from the target population. We assume the following identifiability conditions: (1) independence of the outcome and the population S conditioning on covariates X; (2) the probability of being from the source population conditioning on covariates must be greater than 0 for every x with positive density in the target population(Steingrimsson et al. 2022). These tow fairly strong conditions will allow us to tailor the prediction model and assess its performance in the target population. Given 5 complete imputation datasets from the NHANES data and complete cases from the Framingham data, we will split each of them into training and test sets. To tailor the prediction model $g_{\hat{\beta}}(X)$ for use in the target population, we assume the model $g_{\beta}(X)$ is misspecified in most practical application cases. Then we estimate β using the weighted maximum likelihood estimator, which can be obtained from the inverse-odds weights of being from the source population. Although the inverse-odds weights are not identifiable, we assume, up to unknown proportionally constant, they are equal to the inverse-odds weights in the training set $\frac{Pr(S=0)|X,D_{\text{train}}=1}{Pr(S=1)|X,D_{\text{train}}=1}$(Steingrimsson et al. 2022).

Specifically, we will use 80% of the sex-specific Framingham dataset as training set and 20% as test set, as well as sex-specific imputation datasets. Following the above inverse-odds weights in the training dataset, we firstly combine training set from the Framingham and from each of training imputed 2017-NHANES sets under women and men categories separately, and then fit the logistic model with respect to the population S given covariates mentioned above to get the inverse odds of being from the source population. Since this estimator for the inverse-odds weights is only applicable in the source population, we use the `predict()` function on the training Framingham dataset and take the inverse of exponentiation of predicted outcomes. We tailored the prediction model by adding weights in the `glm()` function with respect to `CVD` and refit the model again to obtain the new estimated β coefficients. Given the tailored prediction model, we plugged into the Framingham test sets and set a threshold of 0.5 to cutoff the binary outcome 0 and 1. To get $\hat{\sigma}(X)$ of the inverse-odds weights in the test set $\frac{Pr(S=0)|X,D_{\text{test}}=1}{Pr(S=1)|X,D_{\text{test}}=1}$, we exponent results from the `predict()` function to the test Framingham data and calculate inverse. Given all those quantities, we estimate the Brier risk scores in the target population following the equation: $\hat{\psi}_{\beta} = \frac{\sum_{i=1}^{n} I(S_i=1,D_{\text{test},i}=1)\hat{\sigma}(X_i)(Y_i-g_{\hat{\beta}}(X_i))^2}{\sum_{i=1}^{n} I(S_i=0,D_{\text{test},i}=1)}$.

Table 3: Estimated Brier Scores in the Target Population

	Men	Women
Composite Fram and Imp 1 Test	0.1677	0.0408
Composite Fram and Imp 2 Test	0.1892	0.0438
Composite Fram and Imp 3 Test	0.1963	0.0342
Composite Fram and Imp 4 Test	0.1898	0.0529
Composite Fram and Imp 5 Test	0.1993	0.0414
Average Estimation for Brier Risk	**0.1885**	**0.0408**

Our findings are relevant to scenarios where prediction models are constructed using training data and subsequently assessed using test data. In these cases, the composite dataset is divided into separate training and test sets to facilitate model development and evaluation. Brier scores, which range from 0 to 1, serve as a measure of predictive accuracy, with 0 denoting perfect accuracy and 1 indicating perfect inaccuracy. Our results reveal that the estimated Brier scores for the NHANES target population closely approach 0 when stratified by gender. Notably, these scores surpass the true scores of 0.1909 for males and 0.1009 for females observed in the source population. This underscores the impressive performance of our customized prediction model in the context of transportability analysis, particularly within the female subgroup.

Simulation Studies

Now we assume that individual level data is not available from the target population and only summary statistics with mean and standard deviation derived from the NHANES dataset are available.

Aim: The simulation study focused on investigating how estimation for the Brier score could be influenced in the simulated target population under different data generation processes for covariates, and examining the number of simulations that could vary estimates.

Data Generation Mechanism:

- We took log transformation for each continuous variable in the source population to ensure normality and got correlation matrix. Then we simulated 3000 observations from a multivariate normal distribution with defined mean and standard deviation values derived from the NHANES-2017 data. As for categorical variables, we fitted logistic models to determine the potential correlations between categorical variables and existing simulated variables.

- We determined the distributions of continuous variables and simulated new individual-level data under certain parameter settings by using `descdist()` function with bootstrapping method and `fitdistr()` function for parameters; Given simulated variables, we still fitted logistic models for categorical variables to understand the association between specified categorical variables and other generated covariates.
- We explored a different distribution of `Age`, a normal distribution, other than the uniform distribution defined in the second data generation method. Then we followed the same procedure of logistic models implemented in the second data generation mechanism to simulate categorical covariates.

Methods: The estimands are described in the estimator for Brier risk in the section of transportability analysis.

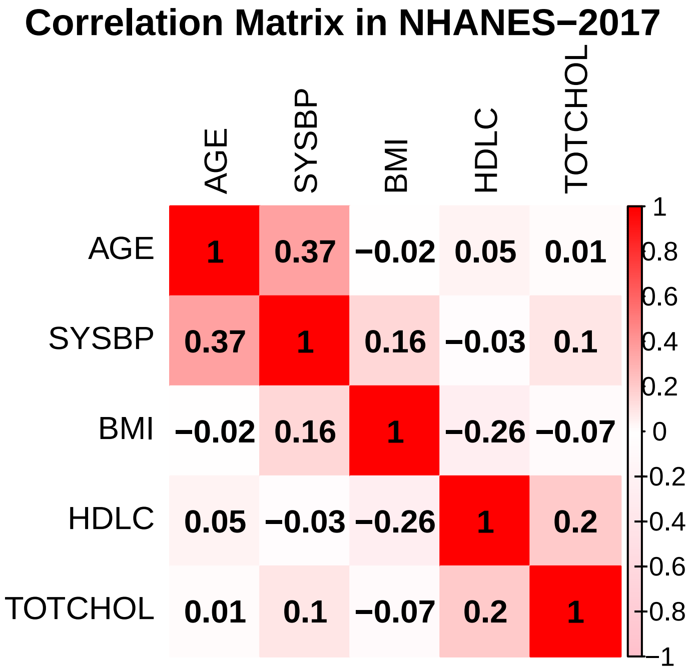
Methods: We split the dataset into training and test sets and applied two sex-specific logistic regression models. Following the inverse-odds weights in the training dataset, we tailored the prediction model to obtain new estimated coefficients and plugged into the test sets to estimate the Brier scores.

Performance Measures: We assessed the average and standard deviation of Brier scores across different numbers of simulations from 1 to 2000, under each simulation case and conducted a direct comparison between the estimations and the original Brier scores from the NHANES-2017 dataset.

The shared parameter is the number of samples to generate (N = 3000) and significance level of 0.05. As for the first data generation method, we took log transformation for each continuous variable to ensure the normality and tested the correlation matrix. Given fixed mean and standard deviation derived from the NHANES data, we simulated from a multivariate normal distribution with defined mean vector and covariance matrix of the continuous variables with the consideration of standard deviation. Except for continuous variable, we also considered about potential correlation between categorical variables, such as use of medication, and continuous variables, such as systolic blood pressure. To get the sense of association, we fitted the logistic model to determine if variables are highly associated with binary outcomes in the Framingham dataset. The first binary variable we examined is `BPMEDS`. The p-value showed that systolic blood pressure plays an important role in predicting the odds of use of medications, thus, we predicted `BPMEDS` in the simulated target population given the logistic regression model with respect to `BPMEDS` given important continuous predictors in the source population. The second categorical variable is `SEX`. We fitted the model with respect to simulated log-transformation of `BMI`, `HDLC`, and `TOTCHOL` and `BPMEDS` to predict the gender category. The third covariate to simulate is `CURSMOKE` given the logistic model including log of `BMI`, log of `AGE` and `SEX` as predictors. The last categorical variable is `DIABETES`. From the insight of important variables correlated to diagnoses of diabetes in the source population, we fitted the logistic model with predictors of systolic blood pressure, age and BMI levels. By continually fitting logistic regression models in terms of categorical covariates, we get primary ideas about potential associations between categorical and continuous variables and simulate binary covariates based on estimated coefficients of significant predictors.

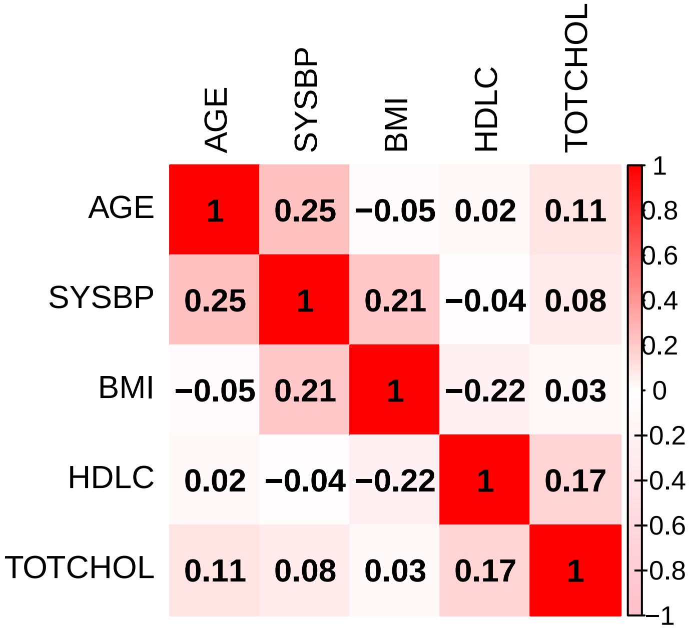
In the second data generation mechanism, we focused on establishing the distributions of continuous covariates within the source population and subsequently simulated new individual-level data based on specific parameter settings. To achieve this, we employed the `descdist()` function, coupled with bootstrapping techniques, to calculate descriptive parameters for empirical distributions of non-censored data, visualizing these through skewness-kurtosis plots. For example, the Cullen and Frey graph revealed that systolic blood pressure closely adhered to a gamma distribution in the source population, as evidenced by the observed data points (in blue) closely aligning with the theoretical gamma distribution dashed line. We then utilized the `fitdistr()` function to estimate the shape and rate parameters, ultimately simulating `SYSBP` according to a gamma distribution with a shape parameter of 40.18 and a rate parameter of 0.29. Following a similar approach, we initially established theoretical distributions for each continuous variable based on Cullen and Frey graphs, subsequently using fixed mean and standard deviation values for the generation of all continuous variables. In summary, `AGE` adheres to a uniform distribution ranging from 30 to 74, while systolic blood pressure SYSBP follows a gamma distribution with the parameters mentioned above. BMI follows a log-normal distribution with a mean of 30.32 on the log scale and a standard deviation of 7.33 on the log scale, HDLC follows a log-normal distribution with a mean of 52.89 on the log scale and a standard deviation of 15.86 on the log scale, and TOTCHOL follows a log-normal distribution with a mean of 192.66 on the log scale and a standard deviation of 41.26 on the log scale. For categorical variables, we adopted a similar procedure to the one employed in the first data generation mechanism. Once again, we fitted logistic regression models to elucidate associations between pre-specified categorical variables and other vital simulated covariates. The inclusion of additional variables in our simulations resulted in increased accuracy for the generation of binary variables of interest.

In the third simulation scenario, we exclusively explored an alternative distribution for AGE, which now adheres to a normal distribution with a mean of 52.41 and a standard deviation of 12.60. All other procedures remained consistent with those performed in the second case. Below, we can find correlation matrices comparing the original target population to the three simulated cases, as well as distribution plots depicting the behavior of each continuous variable under the three simulation scenarios and the true distribution derived from NHANES-2017 data.



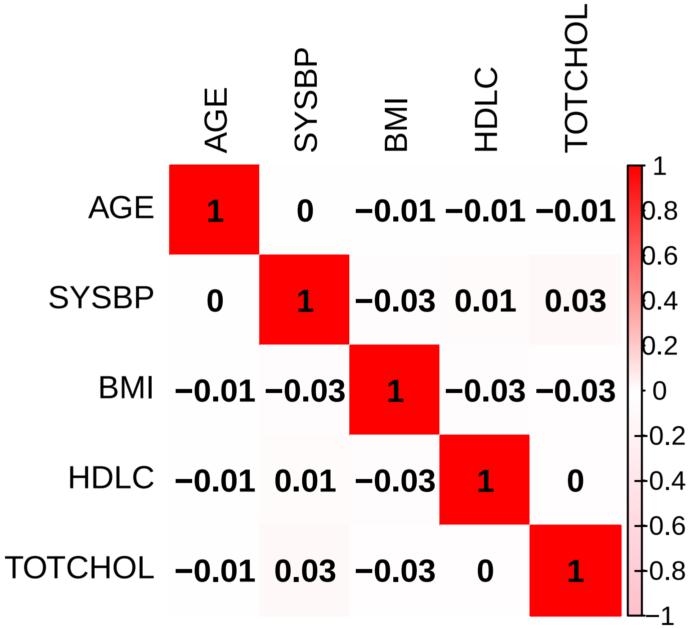
	AGE	SYSBP	BMI	HDLC	TOTCHOL
AGE	1	0.37	−0.02	0.05	0.01
SYSBP	0.37	1	0.16	−0.03	0.1
BMI	−0.02	0.16	1	−0.26	−0.07
HDLC	0.05	−0.03	−0.26	1	0.2
TOTCHOL	0.01	0.1	−0.07	0.2	1

Correlation Matrix in Simulation Case I



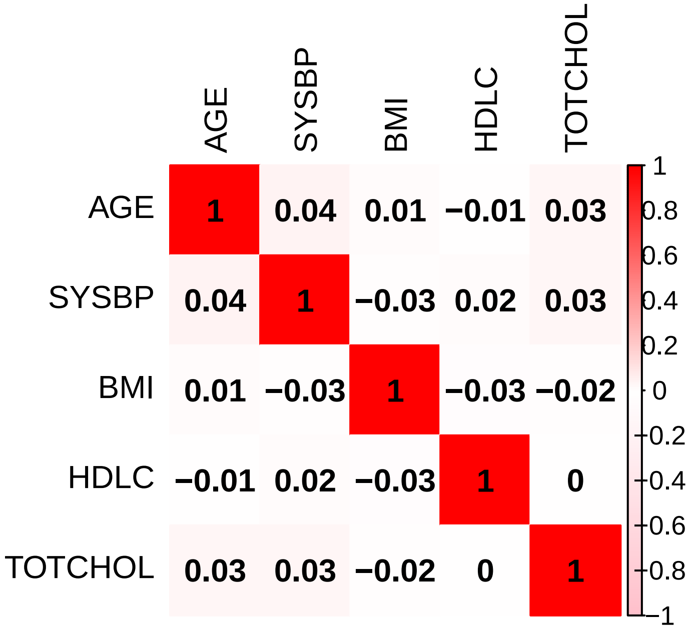
	AGE	SYSBP	BMI	HDLC	TOTCHOL
AGE	1	0.25	−0.05	0.02	0.11
SYSBP	0.25	1	0.21	−0.04	0.08
BMI	−0.05	0.21	1	−0.22	0.03
HDLC	0.02	−0.04	−0.22	1	0.17
TOTCHOL	0.11	0.08	0.03	0.17	1

Correlation Matrix in Simulation Case II



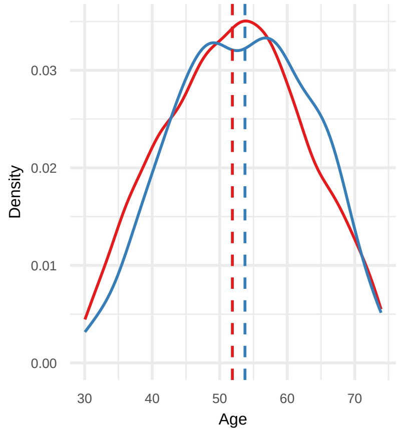
	AGE	SYSBP	BMI	HDLC	TOTCHOL
AGE	1	0	−0.01	−0.01	−0.01
SYSBP	0	1	−0.03	0.01	0.03
BMI	−0.01	−0.03	1	−0.03	−0.03
HDLC	−0.01	0.01	−0.03	1	0
TOTCHOL	−0.01	0.03	−0.03	0	1

Correlation Matrix in Simulation Case III

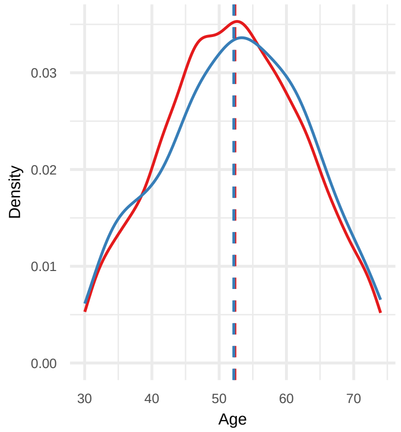


	AGE	SYSBP	BMI	HDLC	TOTCHOL
AGE	1	0.04	0.01	−0.01	0.03
SYSBP	0.04	1	−0.03	0.02	0.03
BMI	0.01	−0.03	1	−0.03	−0.02
HDLC	−0.01	0.02	−0.03	1	0
TOTCHOL	0.03	0.03	−0.02	0	1

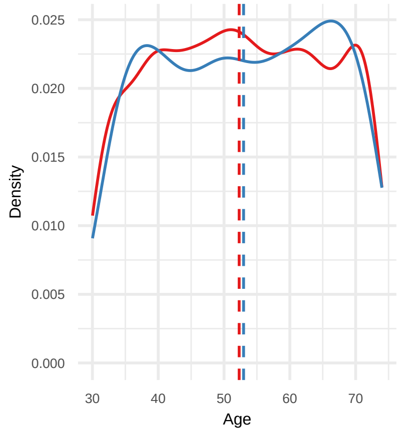
Distribution of Log Transformation of Age



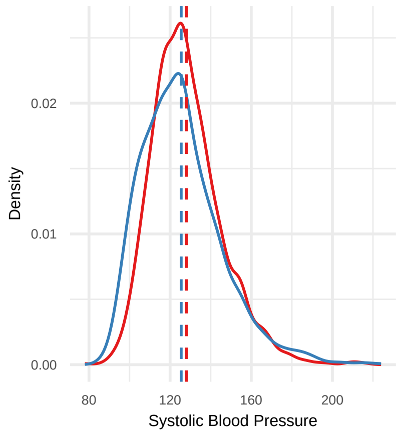
Normal Distribution of Age



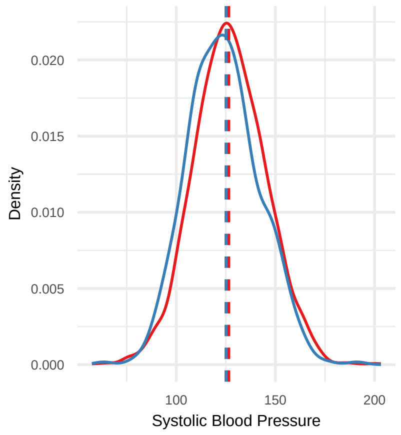
Uniform Distribution of Age



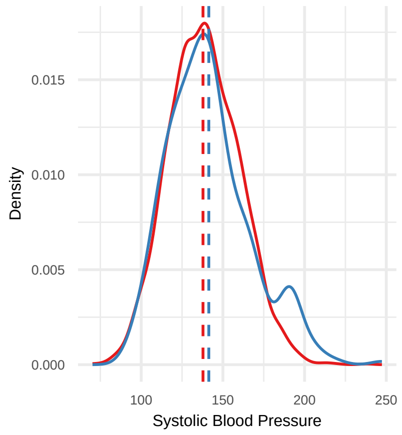
True Distribution of Systolic Blood Pressure in NHANES data

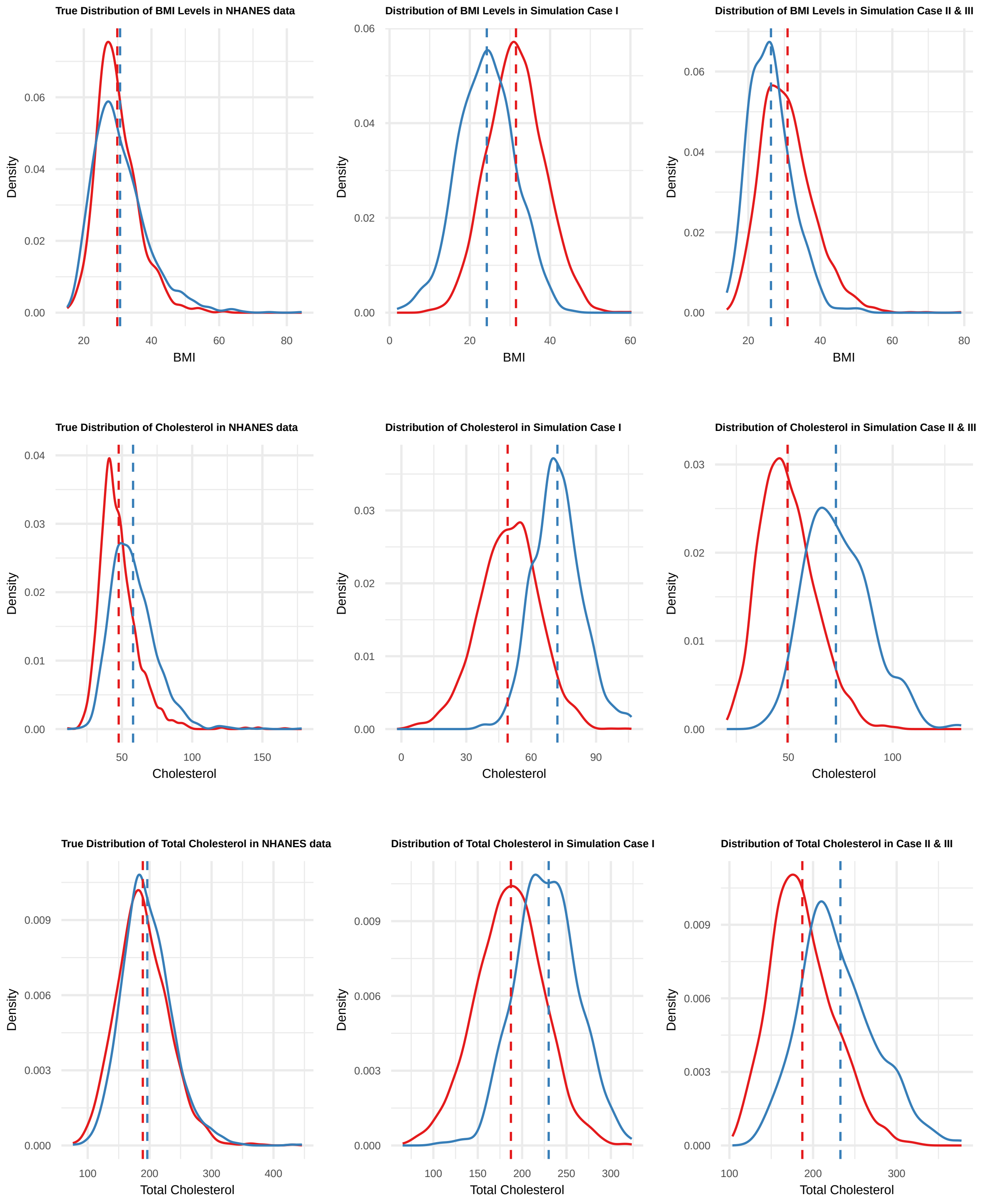


Distribution of Systolic Blood Pressure in Simulation Case I



Distribution of Systolic Blood Pressure Case II & III





True Distribution of BMI Levels in NHANES data

Distribution of BMI Levels in Simulation Case I

Distribution of BMI Levels in Simulation Case II & III

True Distribution of Cholesterol in NHANES data

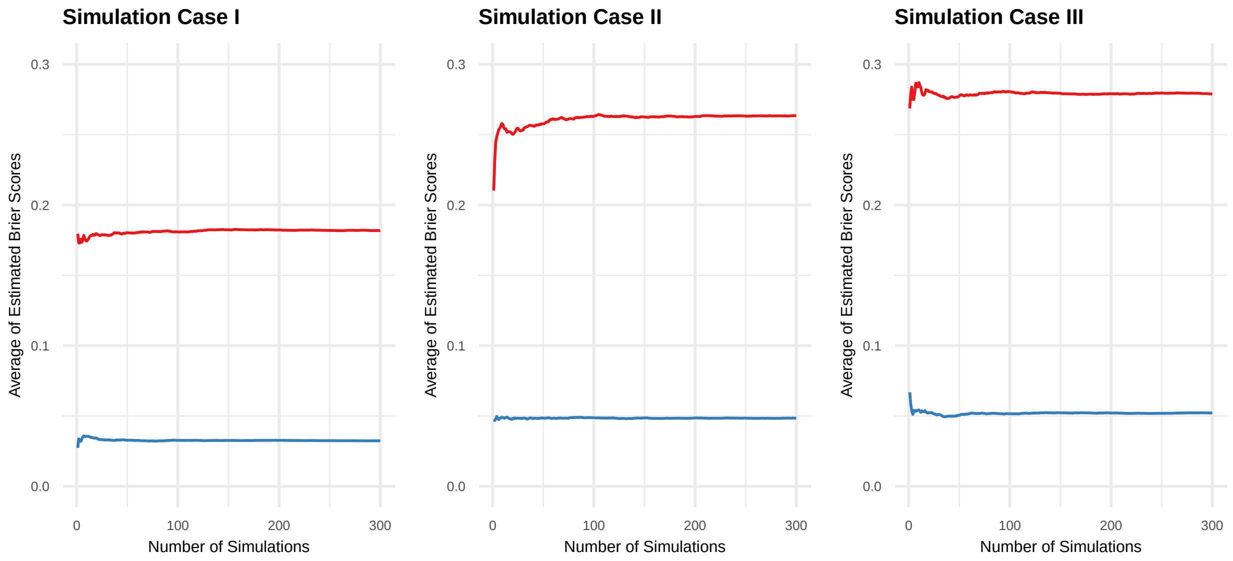
Distribution of Cholesterol in Simulation Case I

Distribution of Cholesterol in Simulation Case II & III

True Distribution of Total Cholesterol in NHANES data

Distribution of Total Cholesterol in Simulation Case I

Distribution of Total Cholesterol in Case II & III



The above plot displays three simulation cases, each showing the average of estimated Brier scores as a function of the number of simulations. In all cases, the average Brier score stabilizes as the number of simulations are greater than 300, indicating the robustness of the simulation process. Case II and Case III show a higher average Brier score in both groups compared to Case I, which could suggest that the model's predictive performance differs across the scenarios. This stability and difference in levels of Brier scores among the cases could reflect the impact of including covariate associations in the simulations, as higher scores indicate a weak performance of transportability in the target set.

Table 4: Estimated Brier Scores in the Simulated Target Population

	MenAvg	MenSD	WomenAvg	WomenSD
Simulation Case I	**0.1813**	**0.0119**	**0.0326**	**0.0046**
Simulation Case II	0.2773	0.0192	0.0526	0.0084
Simulation Case III	0.2649	0.0190	0.0483	0.0081

Table 4 indicates that in Simulation Case I, men had an average Brier score of 0.1813 with a standard deviation (SD) of 0.0119, and women had an average of 0.0326 with an SD of 0.0046. Simulation Case II shows an increase in average Brier scores for both men (0.2773) and women (0.0526), with corresponding increases in SD. Simulation Case III presents average scores slightly lower than Case II but higher than Case I for both genders, indicating misspecification of distributions of variables and the importance of covariate associations in the model's predictive accuracy. For example, the above plots compare the distributions of BMI and cholesterol levels stratified by gender between the NHANES data and three simulation cases, where the red line always represents males and the blue line always represents females. The BMI distributions in the simulation case I deviance a lot from the NHANES data. The same can be observed for the cholesterol and total cholesterol levels, with large deviations in the spreads of the distributions between the true data and the simulations. These plots suggest that the simulations of BMI and cholesterol levels cannot reliably replicate the true distribution of these health-related metrics. Nevertheless, distributions of age and systolic blood pressure in the simulation cases are closely aligned with the NHANES data, indicating that the simulations are accurately reflecting the true data.

Conclusions and Limitations

In our simulation design, the primary objective is to estimate Brier risk scores in the simulated populations using the methodologies outlined in previously published works. Subsequently, we leverage these scores, segmented by gender, to assess the performance of tailored models within our simulated target population. Table 4 presents a breakdown of three Brier scores for each data generation method, categorized by sex. Notably, the Brier scores in the first simulation case fall below the estimations derived from the 2017 NHANES data, suggesting that our simulation closely approximates the true covariate distributions found in the NHANES sample data. However, it is evident that the second and the third data generation methods have higher Brier scores in two groups, as these two methods do not accurately capture covariate associations in the process of simulation. Based on the correlation matrices shown above, we observe that only the first simulation case exhibits a correlation pattern that closely resembles that of the NHANES data, whereas simulation cases II and III demonstrate a notably weaker correlation similarity with the true target population.

The primary limitation of our simulation study arises from the sequential nature of the simulation process. For instance, although we have identified certain categorical variables that play significant roles in predicting specific categorical outcomes, such as gender, the order in which the simulation unfolds prevents us from possessing information about covariates before simulating gender. Consequently, we are constrained in our ability to generate potentially correlated data at this stage. Furthermore, our simulation study does not account for various factors that may vary, such as differing correlation matrices among covariates. These limitations underscore the need for further research to refine and enhance our modeling approaches.

References

D'Agostino, Ralph B., Ramachandran S. Vasan, Michael J. Pencina, Philip A. Wolf, Mark Cobain, Joseph M. Massaro, and William B. Kannel. 2008. "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study." *Circulation* 117 (6): 743–53. https://doi.org/10.1161/circulationaha.107.699579.

National Center for Health Statistics. 2020. "National Health and Nutrition Examination Survey 2017-2018 Data Documentation, Codebook, and Frequencies." Online. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BPQ_J.htm.

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.