Assessing Model Transportability: Performance Evaluation Across Varied Simulated Populations

Monica Colon-Vargas[1], Dr. Alice Paul[1], Dr. Jon Steingrimssoning[1]
[1]Brown University, RI

Overview

The development of predictive models plays a crucial role in healthcare decision-making and risk assessment. However, ensuring the efficacy of these models across diverse populations—known as model transportability—remains a critical challenge. We estimate the Brier score to evaluate the performance in a new target population

Data

Framingham

- Initiated in 1948, the Framingham Heart Study, conducted in Framingham, Massachusetts, is a multigenerational investigation into CVD causes, tracking over 5,000 diverse participants initially.
- Individuals aged 30-62, free from prior heart attacks or strokes, aiming to understand common factors contributing to heart disease.

NHANES

- Conducted in two phases, NHANES comprises thorough home interviews and detailed health examinations.
- Designed to encompass the US household population.

Evaluation Metrics

- Two sex-stratified models were fitted in Framingham dataset.
- We assess model performance using the squared error loss function, aiming to determine and estimate its expectation, known as the mean squared error (MSE) or Brier score (for classification), within this population.
- We use the estimator given in Steingrimssonig, et all defined as follows: $\hat{\varphi}_{\hat{\beta}} = \frac{\sum_{i=1}^{n} I\left(S_i=1,D_{test,i}=1\right)\hat{o}(X_i)(Y_i-g_{\hat{\beta}}(X_i))\text{^}2}{\sum_{i=1}^{n} I\left(S_i=0,D_{test,i}=1\right)}$ where $\hat{o}(X_i)$ is an estimator for the inverse-odds weights in the test set: $\frac{P(S=0|X,D_{test}=1)}{P(S=1|X,D_{test}=1)}$ which is typically obtained with fitting a model for the probability that $S=1$ conditional on the covariates.
- Note: $\hat{\varphi}_{\hat{\beta}}$ is an estimate of the Brier score, not the real Brier score. When the target population lacks the outcome, the real Brier score can not be calculated.

Simulation Framework

- **Aim**: The simulation investigates how altering correlations between variables, based on fixed NHANES statistics, influences Brier score estimation for the target population.
- **Data Generating Mechanism**: This is described in four steps

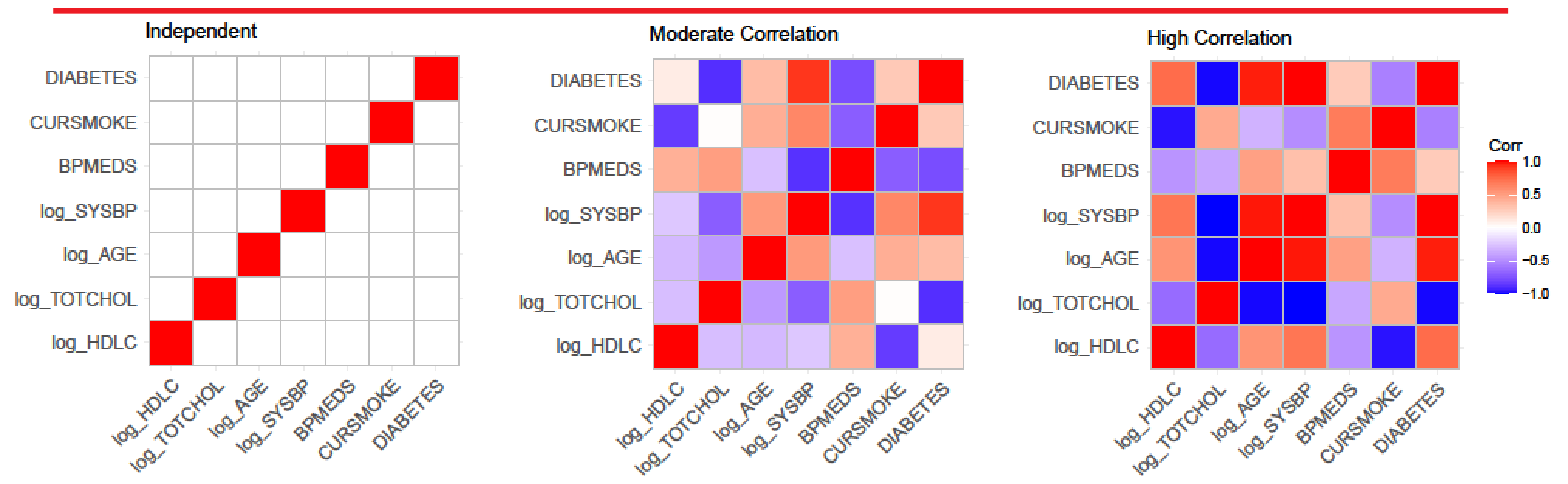
  1. Continuous variables are transformed to logarithmic scale.

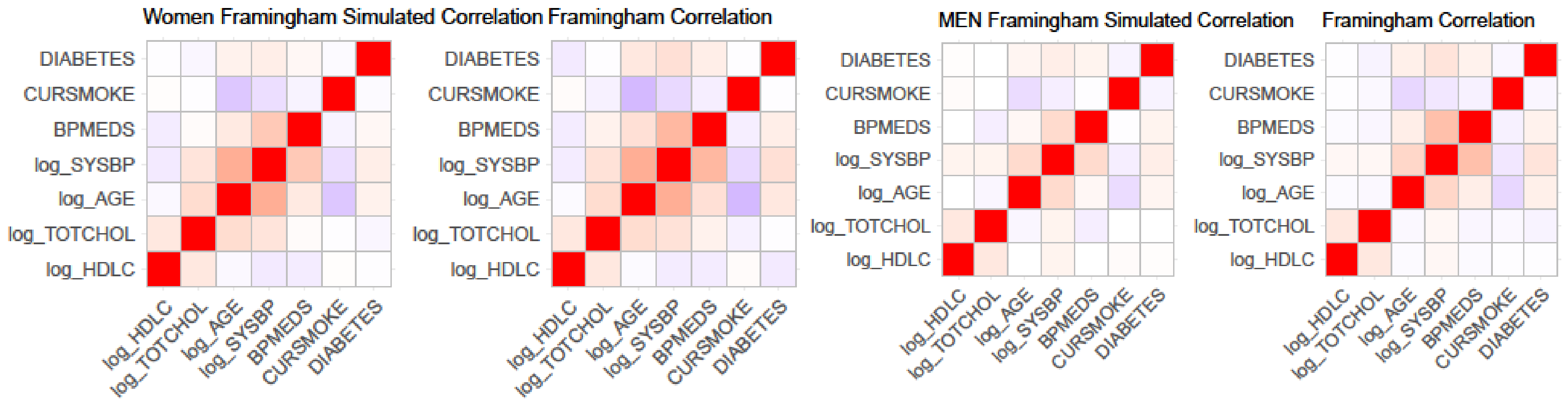
  2. Multivariate random samples of logarithms for these variables are drawn. Samples are drawn with sizes aligned to NHANES dataset: n=3055 for men and n=3544 for women. The multinormal distribution is based on mean and standard deviation parameters of transformed variables, and discrete variable statistics.

  3. The focus is on estimating the Brier score across four correlation settings: independent variables, low, moderate, and high correlation. A similar correlation to the Framingham's data is also considered.

  4. Continuous distributions for discrete variables are addressed using a mean-quantile approach.
- **Estimands**: We use the estimate for the Brier score given in the evaluation metrics.
- **Methods**: Two sex-stratified logistic regression models predicting CVD are tested. The computation of inverse-odds weights relies on a logistic regression model to predict $S=1$.
- **Performance Measures**: We calculate bias between the estimated brier scores on the simulated data and the estimated brier score of the original NHANES dataset.

Scenarios





Results

Women			
Correlation Scenario	**Brier Estimate**	**SD Estimate**	**Bias**
Uncorrelated	0.1158	0.0011	0.0652
Similar	0.1159	0.0010	0.0654
Moderately	0.1151	0.0011	0.0645
High	0.1161	0.0007	0.0655

NHANES Estimated Brier Score: 0.0506
Framingham Brier Score: 0.1160

Men			
Correlation Scenario	**Brier Estimate**	**SD Estimate**	**Bias**
Uncorrelated	0.1918	0.0010	0.0924
Similar	0.1918	0.0010	0.0925
Moderately	0.1916	0.0010	0.0923
High	0.1920	0.0090	0.0926

NHANES Estimated Brier Score: 0.0993
Framingham Brier Score: 0.1919

Conclusion

- Estimated Brier scores in NHANES were lower than those in the source population, potentially due to the relatively better health status of males in NHANES compared to the Framingham dataset.
- Simulated NHANES data exhibited elevated Brier scores for both genders potentially attributed to the simulated datasets' failure to capture the true underlying relationships.
- Despite variations in correlation matrices, the estimates consistently converged, showcasing precision in estimation with low standard errors, indicating stability and reliability in the estimation process.
- The remarkable resemblance between estimated and actual Brier scores from the Framingham dataset across genders suggests the models' potential adaptability and transferability to diverse populations, hinting at their generalizability.

PHP2550: Practical Data Analysis