# Evaluating the Performance of Simulated Datasets in Transportability Analysis

Hannah Eglinton
Brown University, RI

## Overview

**Using a simulation study, we tested whether simulated target population data can be used in lieu of individual-level target population data in the context of transportability analyses.**

## Background

- Risk prediction models are crucial for clinical decision-making but may face challenges when applied to populations with different characteristics.
- Transportability analyses adjust model performance for these differences, but they require individual-level data from the target population, which may not always be available.
- Simulating the target population data from summary statistics may be one way to overcome this challenge.

## Setting

- We fit a cardiovascular disease (CVD) risk model using data from the Framingham Heart Study (FHS), our **source population**.
- The model coefficients included log of HDL cholesterol, log of total cholesterol, log of age, log of systolic blood pressure (SBP) if not treated with BP medication, log of SBP if treated with BP medication, smoking status, and diabetes status.
- We investigated how well this model transported to our **target population**, the 2017-2018 cohort of the National Health and Nutrition Examination Survey (NHANES).

Table 1. Population Characteristics in the FHS and NHANES Cohorts

| **Characteristic** | **FHS**, N = 1,094 | **NHANES**, N = 1,481 |
|---|---|---|
| CVD | 33% | – |
| Age | 60 (8) | 62 (11) |
| Systolic Blood Pressure (mmHg) | 139 (21) | 132 (19) |
| – SBP if Not Treated | 136 (19) | 129 (18) |
| – SBP if Treated | 159 (23) | 135 (18) |
| Diastolic Blood Pressure (mmHg) | 82 (11) | 75 (13) |
| HDL Cholesterol (mg/dL) | 44 (13) | 49 (14) |
| Total Cholesterol (mm/dL) | 226 (41) | 187 (42) |
| BMI | 26.2 (3.5) | 29.2 (6.0) |
| Cigarette Smoker | 39% | 20% |
| Diabetic | 8.8% | 22% |
| Uses Anti-hypertensive Medication | 11% | 41% |

[1] %; Mean (SD)

## Transportability Analysis

- The **Brier score** in the target population was estimated as follows, where $S_i = 1$ for individuals in FHS, $S_i = 0$ for individuals in NHANES, and $g(X_i)$ is the CVD risk as predicted by the FHS model.

$$\hat{\psi} = \frac{\sum_{i=1}^{n} I(S_i = 1)\hat{o}(X_i)(Y_i - g(X_i))^2}{\sum_{i=1}^{n} I(S_i = 0)}, \text{ where } \hat{o}(X_i) = \frac{Pr[S = 0|X]}{Pr[S = 1|X]}$$

- The **AUC** in the target population was estimated as follows:

$$\hat{\tau} = \frac{\sum_{i \neq j} w(X_i, X_j) I(g(X_i) > g(X_j), Y_i = 1, Y_j = 0, S_i = 1, S_j = 1)}{\sum_{i \neq j} I(Y_i = 1, Y_j = 0, S_i = 1, S_j = 1))},$$

$$\text{where } w(X_i, X_j) = \frac{P[S = 0|X_i]P[S = 0|X_j]}{P[S = 1|X_i]P[S = 1|X_j]}$$

## Simulation Study

- The aim of our simulation design was to evaluate the covariate association assumptions needed to replicate individual-level NHANES data in transportability analyses.
- Operating as if only the NHANES summary data in Table 1 were available, we simulated the model covariates based solely on these values.
- We assumed that HDL cholesterol, total cholesterol, and SBP followed a normal distribution, that age followed a uniform distribution, and that BP medication, smoking, and diabetes followed Bernoulli distributions.
- The **correlation between HDL cholesterol and total cholesterol ($\rho$)** was varied from 0.0 to 0.9.
- The **association between age and BP medication ($\gamma$)** was also varied from 0.0 to 0.9.
- The estimands of interest were the Brier score and AUC as estimated through transportability analyses using individual-level NHANES data.
- We generated $n_{sim} = 100$ datasets for each of the 100 situations (all combinations of the 10 values for $\rho$ and 10 values for $\gamma$).
- We assessed the **relative bias** of our Brier score and AUC estimates derived from simulated data ($\hat{\theta}$) compared to the Brier score and AUC estimands determined using the individual-level data ($\theta$).

$$\text{Relative Bias} = \frac{1}{n_{sim}} \sum_{i=1}^{n} \frac{\hat{\theta}_i - \theta}{\theta}$$

## Results

Figure 1. Brier Score Bias for Men (solid) and Women (dashed)

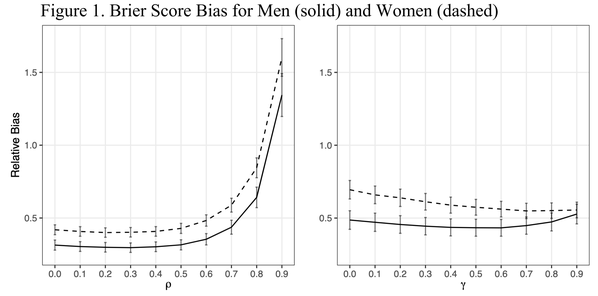


Figure 2. AUC Bias for Men (solid) and Women (dashed)

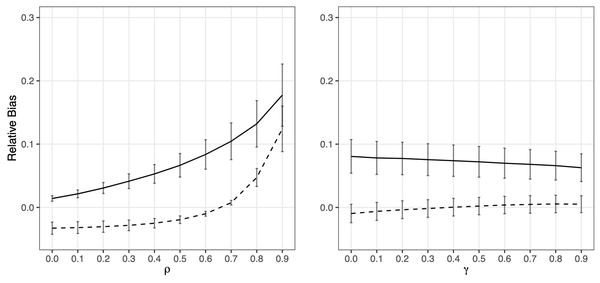


Table 1. Comparison of Association Assumptions

| | $\theta$ | $\hat{\theta}^{a}_{\text{NHANES}}$ | $\hat{\theta}^{b}_{\text{FHS}}$ | $\hat{\theta}^{c}_{\text{none}}$ |
|---|---|---|---|---|
| **Brier Score** | | | | |
| Males | 0.118 | 0.151 (28.0%) | 0.153 (29.7%) | 0.156 (32.2%) |
| Females | 0.064 | 0.090 (40.6%) | 0.091 (42.2%) | 0.096 (50.0%) |
| **AUC** | | | | |
| Males | 0.758 | 0.782 (3.17%) | 0.779 (2.77%) | 0.773 (1.98%) |
| Females | 0.824 | 0.798 (3.16%) | 0.795 (3.52%) | 0.790 (4.13%) |

[a] Male data simulated with $\rho = 0.18$ and $\gamma = 0.25$; Female data simulated with $\rho = 0.21$ and $\gamma = 0.30$.
[b] Male data simulated with $\rho = 0.14$ and $\gamma = 0.14$; Female data simulated with $\rho = 0.12$ and $\gamma = 0.20$.
[c] Male and female data simulated with $\rho = 0.0$ and $\gamma = 0.0$.

## Conclusions

- The low relative biases observed suggest that using simulated data to conduct transportability analysis is a valid way to estimate the Brier score and AUC in a target population when individual-level data are not available.
- Our results suggest that researchers can either assume no associations between covariates or use the associations observed in the source population.
- These assumptions make the simulations simple to implement, since they don't require guessing the exact target population associations.