Evaluating the Performance of Simulated Datasets in Transportability Analysis

Hannah Eglinton

2023-11-20

Abstract

Background: Risk prediction models, such as the Framingham ATP-III model for cardiovascular disease, are crucial for clinical decision-making but may face challenges when applied to populations with different characteristics. Transportability analysis techniques can 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.

Methods: Simulated target population data was generated using NHANES summary statistics and a variety of correlation parameters. Transportability analyses were conducted on each simulated dataset to estimate the model Brier score and AUC. The bias of these estimates compared to the estimated Brier score and AUC derived using individual-level NHANES data was calculated for each correlation scenario.

Results: Transportability analyses using simulated data sets resulted in relative biases between -0.05 and 1.5 in comparison to the estimates derived from individual-level data. While adjusting the extent of association between particular simulated variables could improve performance, the simulations in which assumed no associations were not substantially different.

Conclusions: The low relative biases observed suggest that using simulated data to conduct transportability analysis is a valid way to estimate Brier scores and AUC in a target population when individual-level data is not available. Further, our results suggest that researchers can assume no association between covariates when simulating the target data, making the target data simulation simple to implement.

Introduction

Risk prediction models can play a critical role in clinical decision-making, offering tools to identify individuals at heightened risk for a variety of adverse events. However, the successful deployment of these models relies on their ability to generalize across diverse populations. A common challenge arises when models developed on one population are applied to a population with different characteristics, which can lead to lowered model performance in these novel settings. The Framingham ATP-III model [1], a widely used predictor of 10-year cardiovascular disease (CVD) risk, serves as a fitting example. Initially constructed using data from the Framingham Heart Study (FHS), predominantly from white participants, its generalizability to the general population has been questioned.

Transportability analysis offers techniques to weight the performance metrics of models to account for differences in the data distribution between the source population and the target population [2,3]. Further, these techniques can be used even when the outcome measure was not collected in the target population. However, transportability analysis methods do require individual-level data on the target population. Critically, individual-level data are not always available for the target population of interest, as many researchers have access only to the summary statistics of the target population.

Simulating the individual-level data from these summary statistics may be one way to apply transportability

analyses in this situation. To demonstrate this, our study fit a CVD risk model on FHS data and then conducted transportability analyses using a simulated target population from the National Health and Nutrition Examination Survey (NHANES), a nationally representative dataset that does not contain long-term outcome data on CVD. Our study evaluated how well simulated data can replicate individual-level target population data in the context of transportability analysis, specifically investigating how important correlation assumptions are when simulating the data. We assessed the bias of Brier score and AUC estimates derived through transportability analysis using simulated data in comparison to those derived using the individual-level NHANES data.

Methods

Data Sources

The Framingham Heart Study (FHS) began in 1948 as a prospective study on cardiovascular disease (CVD) in a population residing in Framingham, Massachusetts. Since then, over 14,000 people from three generations have participated as subjects in FHS, including the original participants, their children, and their grandchildren. The study comprises a repository of demographic, clinical, and lifestyle data over multiple examination cycles, allowing researchers to investigate risk factors of cardiovascular disease. Eligible participants were between 30 and 62 years at the time of the first examination and did not have cardiovascular disease at these baseline examinations [4].

For this study, a subset of data from the FHS was accessed using the `riskCommunicator` package in R, which included laboratory, clinic, questionnaire, and event data on 4,434 participants. The subset included data from three examination periods, spanning approximately six years each, from around 1956 to 1968. Each participant was followed for 24 years to assess the occurrence of CVD, which included angina pectoris, myocardial infarction, atherothrombotic infarction, and cerebral hemorrhage.

NHANES is a comprehensive and nationally representative survey conducted by the Centers for Disease Control and Prevention and the National Center for Health Statistics. The FHS and NHANES populations differ in a number of ways. NHANES employs a nationally representative sampling strategy, encompassing diverse demographic groups across the United States. In contrast, the FHS originated with a geographically limited cohort, comprising of residents of Framingham, Massachusetts. NHANES captures a more diverse range of demographic backgrounds, including various ethnicities, socioeconomic statuses, and geographic regions. The FHS, especially in its earlier phases, primarily consisted of a white population. NHANES includes representation from urban, suburban, and rural areas across the country while the FHS represents only a single community.

Further, NHANES utilizes a cross-sectional design, while the FHS is a longitudinal study that follows individual participants over multiple examination cycles. Finally, the two cohorts differ on eligibility criteria. While FHS excluded participants with baseline cardiovascular disease, NHANES includes subjects with pre-existing cardiovascular conditions. While FHS required participants to be aged 30 to 62 years at entry, NHANES targets the entire population aged two months and older.

Data Preprocessing

Framingham Heart Study

First, variables used in the Framingham ATP-III model [1] were selected from the FHS data. Specifically, these variables included CVD, sex, age, total cholesterol (mg/dL), high density lipoprotein (HDL) cholesterol (mg/dL), systolic blood pressure (SBP, mmHg), current cigarette smoking status, diabetes status, and anti-hypertensive medication status. CVD included myocardial infarction, fatal coronary heart disease, atherothrombotic infarction, cerebral embolism, intracerebral or subarachnoid hemorrage, and fatal cerebrocascular disease.

The Framingham ATP-III model included separate SBP covariates depending on whether the subject was treated with anti-hypertensive medication. We reproduced these covariates by reformatting the SBP into two

separate covariates. One variable contained the SBP if the subject was *not* on anti-hypertensive medication and was coded as 0 if they were on medication. The second variable contained the SBP if the subject *was* on anti-hypertensive mediation and was coded as 0 if they were not.

Following the methods of the Framingham ATP-III model development, only observations with nonmissing covariate data were included [1]. Since cholesterol was only measured at the third examination period, only this timepoint was included for each subject. In total, 2,578 participants had complete covariate data at the third examination period. The ages of the participants in this period ranged from 44 to 81 years.

We aimed to examine a model that predicted 15-year CVD risk. Therefore, we removed participants that had a time-to-CVD-event of less than 15 years but who were labeled as having no CVD. In other words, censored data were removed, leaving 2,539 participants. The final FHS population characteristics are reported in Tables 2 and 3 for men and women, respectively.

NHANES

The variables selected from the FHS data were selected from the 2017-2018 NHANES dataset. Notably, NHANES is cross-sectional and does not contain long-term outcome information so a CVD outcome variable was not included.

To match the FHS eligibility criteria, we filtered the NHANES data for participants between 44 and 81 years of age. FHS also excluded all individuals with CVD at baseline, namely myocardial infarction, fatal coronary heart disease, atherothrombotic infarction, cerebral embolism, intracerebral or subarachnoid hemorrage, and fatal cerebrocascular disease. We were able to exclude NHANES participants with a history of myocardial infarction and coronary heart disease, though data on the other types of CVD were not collected by NHANES. In total, the final NHANES dataset used in this study included 3,176 participants. The study population characteristics are reported in Tables 2 and 3 for men and women, respectively.

The NHANES dataset exhibited varying degrees of missing data across the different variables used the Framingham ATP-III model, with sex, age, and smoking having complete data, while variables like blood pressure medication, HDL cholesterol, total cholesterol, and systolic blood pressure display notable percentages of missing values (Table 1). This pattern suggests that missingness was primarily influenced by whether the individual attended a laboratory test as part of the NHANES survey. It is unreasonable to assume that missing the laboratory test is completely at random, so we cannot simply exclude patients with missing data.

Table 1: Prevalence of Missing Data in NHANES Cohort

	Percent Missing
Sex	0%
Age	0%
Smoking	0%
Diabetes	0.1%
BP Medication	6.2%
HDL Cholesterol	10.9%
Total Cholesterol	10.9%
Systolic Blood Pressure	16.6%

Instead, the missing data pattern suggests a potential missing at random (MAR) mechanism, especially for variables associated with laboratory measurements. It is reasonable to assume that an individual's likelihood of attending the laboratory portion of the NHANES survey was influenced by factors such as their age, sex, socioeconomic status, health conditions, and lifestyle. Therefore, we used multiple imputation to address the missing data in this dataset by first pulling these relevant factors from the NHANES database. Specifically, in addition to the variables used in Framingham ATP-III, we also included family income, race, and BMI in the multiple imputation model.

Multiple imputation was performed using the `mice` package in R. The MICE algorithm iteratively imputes missing values for each variable based on the observed values of other variables in the dataset. Continuous

variables were imputed using predictive mean matching and categorical variables were imputed using logistic regression. We created five imputed datasets to ensure robustness in our imputation process.

Transportability Analysis

We first fit CVD risk models using the FHS cohort. Separate logistic regression models were fit on the male and female FHS populations. The outcome of our models was whether a CVD event occured within 15 years from baseline. We used the same covariates as were included in the Framingham ATP-III model: log of HDL cholesterol, log of total cholesterol, log of age, log of SBP if not treated, log of SBP if treated, smoking, and diabetes.

Our transportability analyses used the following estimator for the Brier score [2]:

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

Here, $S_i = 0$ indicates that subject i is in NHANES and $S_i = 1$ indicates the subject i is in FHS. X_i are the covariates sex, total cholesterol, age, blood pressure, smoking, diabetes, anti-hypertensive medication, and HDL cholesterol. The outputs to the function $g(X_i)$ are the predicted probabilities of CVD using the model fit on Framingham data.

We used the following estimator for the AUC [3]:

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

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

Simulation Design

Aim: The aim of our simulation design was evaluate the conditions under which simulated data can replicate individual-level NHANES data in transportability analyses.

Data generation: Operating as if only the NHANES summary data from Table 2 and Table 3 were available, we simulated the model covariates based solely on these values.

HDL cholesterol and total cholesterol were generated from the following multivariate normal distribution:

$$N(\mu, \Sigma), \text{ where } \mu = \begin{pmatrix} \mu_{HDL} \\ \mu_{TOT} \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma^2_{HDL} & \rho\sigma_{HDL}\sigma_{TOT} \\ \rho\sigma_{HDL}\sigma_{TOT} & \sigma^2_{TOT} \end{pmatrix}$$

For the male data generation, $\mu_{HDL} = 49$, $\sigma_{HDL} = 14$, $\mu_{TOT} = 187$, and $\sigma_{TOT} = 42$ (Table 2). For the female data generation, $\mu_{HDL} = 59$, $\sigma_{HDL} = 16$, $\mu_{TOT} = 200$, and $\sigma_{TOT} = 41$ (Table 3). The ρ parameter was varied from 0 to 0.9 by 0.1. The use of a multivariate normal distribution reflects the likely correlation between the two types of cholesterol.

Since anti-hypertensive medications are more commonly prescribed in older individuals [5], we included a parameter that represented the association between age and anti-hypertensive medication. We assumed that age followed a $Unif(a, b)$ distribution with $a = 42$ and $b = 82$ for both males and females. Under this distribution, the expected value for age is 62 and the standard deviation is 11.5, which corresponds well to both the observed summary statistics and the known eligibility criteria. We assumed that anti-hypertensive medication followed a $Bernoulli(p)$ distribution where $p = 0.41$ for males and $p = 0.46$ for females.

To achieve an association between age and anti-hypertensive medication, we first generated two variables from the following multivariate normal distribution:

$$N(\mu, \Sigma), \text{ where } \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$$

To create the age covariate, the first generated variable was transformed from a normal distribution to a uniform distribution using the cumulative distribution function of the normal distribution. This transformed variable (now ranging from 0 to 1) was scaled by 40 and shifted by 42 to achieve the desired uniform distribution from 42 to 82.

To create the anti-hypertensive medication covariate, the second generated variable was transformed from a normal distribution to a $Bernoulli(p)$ distribution by coding each value greater than $(1-p)$ as 1 and each value less than $(1-p)$ as 0, where p was the proportion of individuals in the population who were on medication (0.41 for males and 0.46 for females). This created a binary variable that was related to age (with the strength of the association parameterized by γ) and matched the proportions observed in Tables 2 and 3.

For each simulated individual, two SBP variables were generated. A variable reflecting SBP if not treated with blood pressure medications was generated from $N(129, 18)$ for males and from $N(128, 19)$ for females. A second variable that reflected SBP if treated with blood pressure medications was generated from $N(135, 18)$ for males and from $N(139, 20)$ for females. For individuals simulated to be on anti-hypertensive medication, the "SBP if not treated" variable was dropped to zero. Likewise, for individuals simulated to *not* be on anti-hypertensive medication, the "SBP if treated" variable was dropped to zero.

Smoking and diabetes status were each simulated from independent $Bernoulli(p)$ distributions. For the smoking variable, $p = 0.20$ for males and $p = 0.12$ for females. For the diabetes variable, $p = 0.22$ for males and $p = 0.19$ for females. We made the assumption that smoking and diabetes were not related to any variables in the model. While likely not the case in the real data, we did not feel that including these associations would be sufficiently justified if only the summary data were available.

Estimands: The estimands of interest were the Brier score and AUC as estimated through transportability analysis using individual-level NHANES data.

Methods: After setting a randomization seed of 1, we generated data for both males and females under 100 situations (10 possible values for ρ and 10 possible values for γ). For each situation, we simulated $n_{sim} = 100$ datasets. We determined that $n_{sim} = 100$ was sufficient to get a Monte Carlo estimate of bias less than 0.005 based on an initial small simulation run that showed that the standard deviation of our estimands to be less than 0.02.

$$\text{Monte Carlo SE(Bias)} = \sqrt{Var(\hat{\theta})/n_{sim}} = \sqrt{\frac{0.02^2}{100}} = 0.002$$

For each simulated data set, we calculated the estimated Brier score and estimated AUC using the transportability analysis equations defined above using the models fit on FHS data.

Performance measures: We assessed the relative bias of our Brier score and AUC estimates derived from simuluated data ($\hat{\theta}$) compared to the Brier score and AUC estimands determined using the individual-level data (θ). We calculated relative bias instead of bias because men and women have different values for θ, so relative bias was required to compare the performance of the male and female simulations.

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

Results

Population Characteristics and Risk Model

A summary of the population characteristics of men in the FHS and NHANES cohorts is reported in Table 2 and a summary of the population characteristics of women in the FHS and NHANES cohorts is reported

in Table 3. The NHANES cohort generally had lower blood pressure, higher HDL cholesterol, and lower total cholesterol on average. Additionally, the NHANES cohort had a lower prevalence of cigarette smoking, a higher prevalence of diabetes, and a higher prevalence of anti-hypertensive medication than the FHS population. These differences in population characteristics reflect the need for transportability analysis, as we cannot assume that the model fit on the FHS population will have the same performance when used on the NHANES population.

Table 2: Characteristics of Men 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)		

Table 3: Characteristics of Women in the FHS and NHANES Cohorts

Characteristic	**FHS**, N = 1,445	**NHANES**, N = 1,695
CVD	17%	–
Age	61 (8)	62 (11)
Systolic Blood Pressure (mmHg)	140 (24)	133 (21)
– SBP if Not Treated	136 (22)	128 (19)
– SBP if Treated	159 (21)	139 (20)
Diastolic Blood Pressure (mmHg)	80 (11)	72 (15)
HDL Cholesterol (mg/dL)	53 (16)	59 (16)
Total Cholesterol (mm/dL)	246 (46)	200 (41)
BMI	26 (4)	30 (8)
Cigarette Smoker	31%	12%
Diabetic	6.6%	19%
Uses Anti-hypertensive Medication	18%	46%
[1] %; Mean (SD)		

The coefficients and odds ratios from the logistic regression models fit on the FHS data are reported in Tables 4 and 5 for males and females, respectively.

Table 4: Coefficients and Odds Ratios for Male Risk Model

	Estimate	p-value	OR	95% CI
Intercept	-32.96			
Log of HDL cholesterol	-0.66	0.003	0.52	(0.33 - 0.8)
Log of total cholesterol	1.03	0.008	2.8	(1.31 - 6)
Log of age	4.79	<0.001	120.66	(40.91 - 355.82)
Log of SBP if not treated	1.87	<0.001	6.51	(2.36 - 18)
Log of SBP if treated	1.95	<0.001	7.01	(2.6 - 18.9)
Smoking	0.26	0.07	1.3	(0.98 - 1.73)
Diabetes	0.76	0.001	2.13	(1.35 - 3.36)

Table 5: Coefficients and Odds Ratios for Female Risk Model

	Estimate	p-value	OR	95% CI
Intercept	-35.15			
Log of HDL cholesterol	-1.22	<0.001	0.29	(0.18 - 0.49)
Log of total cholesterol	0.90	0.04	2.45	(1.04 - 5.75)
Log of age	5.09	<0.001	162.24	(42.66 - 616.99)
Log of SBP if not treated	2.42	<0.001	11.24	(3.84 - 32.9)
Log of SBP if treated	2.49	<0.001	12.01	(4.19 - 34.44)
Smoking	0.52	0.005	1.68	(1.17 - 2.41)
Diabetes	1.09	<0.001	2.96	(1.84 - 4.78)

When applying the male risk model to the source population, the estimated Brier score was 0.192 and the estimated AUC was 0.712. Using the female model on the female source population, the estimated Brier score was 0.116 and the estimated AUC was 0.774. However, note that these metrics were computed on the model's training data so likely overestimate the model's performance.

Transportability Analysis Using Individual-Level Data

Using the weighted estimators, the mean estimated Brier score across the five imputed target male populations was 0.118 and the mean estimated AUC was 0.758. The mean estimated Brier score across the imputed target NHANES female populations was 0.064 and the mean estimated AUC was 0.824. Unexpectedly, these metrics suggest that the model would perform better in the NHANES population than the FHS population that it was trained on.

These weighted estimators derived using the individual-level NHANES data are the estimands of interest in our simulation study.

Transportability Analysis Using Simulated Data

Figure 1 plots the average relative bias of the estimated Brier scores derived from simulated data in comparison to the estimands of $\theta = 0.192$ for men and $\theta = 0.116$ for women. The error bars represent the Monte Carlo standard error for relative bias. Data are averaged over each possible ρ value or each possible γ value. Recall that ρ is the correlation between HDL cholesterol and total cholesterol, while γ modulates the relationship between age and anti-hypertensive medication.

Overall, lower values of ρ performed better in both the male and female simulations. Calues of ρ over 0.6 caused the relative bias to increase substantially. The value of γ didn't have as much of an effect on relative bias, though values closer to 0.5 performed best in the male simulations and values closer to 0.7 performed best in the female simulations. In all cases, the average relative bias when $\rho = 0$ was within one standard error of the optimal ρ value and the bias was $\gamma = 0$ was within one standard error of the optimal γ value.

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

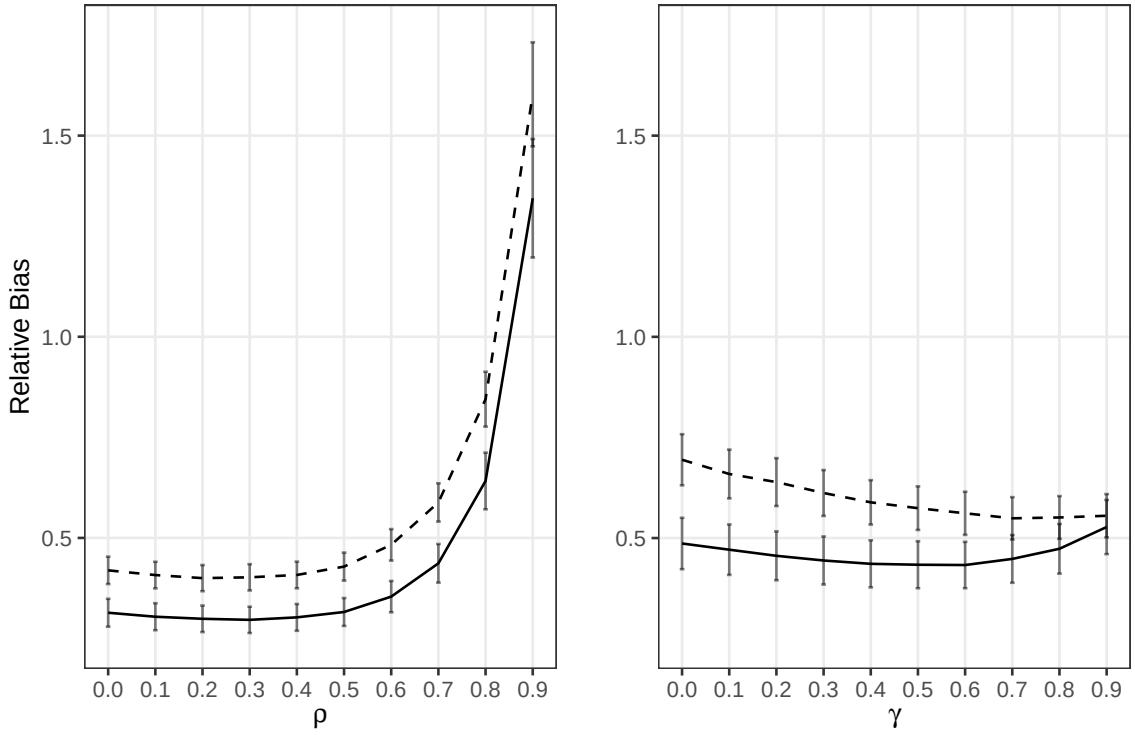

Figure 2. AUC Bias for Men (solid) and Women (dashed)

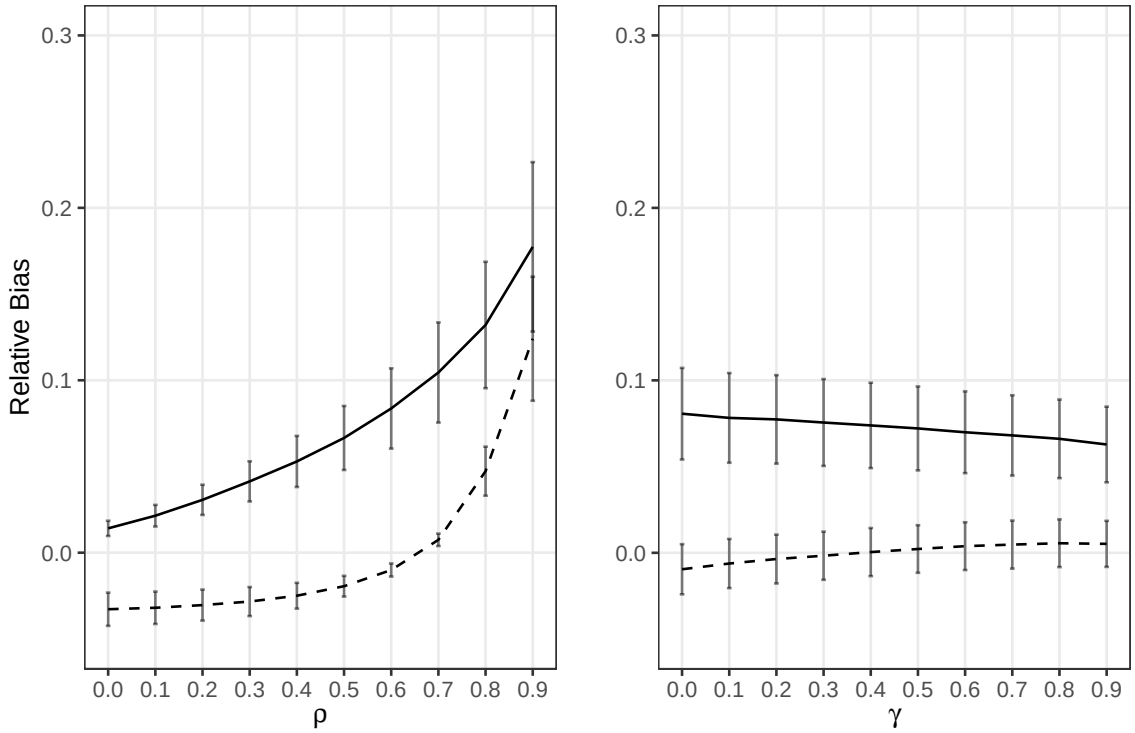

Figure 2 plots the average relative bias of the estimated AUC derived from simulated data in comparison to the estimands of $\theta = 0.758$ for men and $\theta = 0.824$ for women. For both the male and female simulations, the relative bias of AUC increased as ρ increased. While the male simulation always overestimated AUC, the female simulation underestimated AUC when $\rho < 0.7$. Similar to the Brier score results, the relative bias from all γ values were within one standard error of each other, indicating that the γ parameter didn't have a large effect on AUC. However, the AUC relative bias did decrease slightly as γ increased in the male simulation and increased slightly as γ increased in the female simulation.

Table 6 compares the Brier score and AUC estimates when no association assumptions were made ($\rho = 0$, $\gamma = 0$) to the estimates when the optimal assumptions were made. Although the optimal assumptions resulted in estimates closer to θ, the optimal values for ρ and γ were not consistent across estimands and situations and would be difficult to guess based on only summary statistics. On the other hand, the $\hat{\theta}_{\text{none}}$ estimates require no assumptions and are not substantially different that the $\hat{\theta}_{\text{opt}}$ estimates.

Table 6: Comparison of Association Assumptions

	θ	$\hat{\theta}_{\text{none}}$	$\hat{\theta}_{\text{opt}}$
Brier Score			
Males	0.118	0.156	0.149 ($\rho = 0.3$, $\gamma = 0.5$)
Females	0.064	0.096	0.086 ($\rho = 0.2$, $\gamma = 0.9$)
AUC			
Males	0.758	0.773	0.762 ($\rho = 0.0$, $\gamma = 0.9$)
Females	0.824	0.790	0.824 ($\rho = 0.7$, $\gamma = 0.1$)

Discussion

Transportability analysis methods can address the challenge of transporting a model from its source population to a target population without measured outcome data. However, these methods rely on the availability of individual-level data on the target population. This study investigated the bias introduced in transportability analyses when simulating the target population from summary statistics. In particular, we ran a simulation study to investigate how precise association assumptions must be when generating potentially correlated covariates.

Using transportability estimates for Brier score and AUC, we found that the relative bias of simulated data compared to individual-level data could be lower than 0.1. This suggests that using simulated data to conduct transportability analysis is a valid way to estimate Brier scores and AUC in a target population when individual-level data is not available. Further, we determined that while our simulation identified optimal correlation values between simulated covariates, its performance wasn't markedly improved over the simulations using no correlation between covariates. This finding suggests that researchers can assume no association between covariates when simulating the target data, making the method easier to implement.

Our simulation study has a number of limitations. First, we only set associations between the HDL cholesterol and total cholesterol and between age and anti-hypertensive medication. Therefore, we did not investigate all possible correlations between covariates in the model, some of which could potential alter performance. Second, our study looked only at the transportability of a single risk model from one source population (FHS) to one target population (NHANES). Our simulation results are not necessarily generalizable to all transportability analyses. Third, we were not able to exactly replicate the FHS eligibility criteria in the NHANES cohort due to the history of certain CVD conditions not being collected during the NHANES survey.

Finally, our study did not include a train/test split of the data, so the Brier score and AUC for the source population should be interpreted with caution. Since the purpose of our study was to compare transportability analysis results of simulated versus individual-level data, we were not focused on the performance of the

model on the source population or on comparing the transporability analysis results on the target population to the performance of the model on the source population.

In conclusion, our study contributes to understanding how simulated data could be used when transporting a prediction model to a target population with different characteristics. Future research could explore alternative simulation approaches and consider a broader range of variables to enhance the applicability of risk prediction models in varied settings.

References

[1] D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General Cardiovascular Risk Profile for use in primary care. Circulation, 117(6), 743–753. https://doi.org/10.1161/circulationaha.107.699579

[2] Steingrimsson, J. A., Gatsonis, C., Li, B., & Dahabreh, I. J. (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

[3] Li, B., Gatsonis, C., Dahabreh, I. J., & Steingrimsson, J. A. (2022). Estimating the area under the ROC curve when transporting a prediction model to a target population. Biometrics, 79(3), 2382–2393. https://doi.org/10.1111/biom.13796

[4] Andersson, C., Johnson, A. D., Benjamin, E. J., Levy, D., & Vasan, R. S. (2019). 70-year legacy of the Framingham Heart Study. Nature Reviews Cardiology, 16(11), 687–698. https://doi.org/10.1038/s41569-019-0202-5

[5] Singh, J. N., Nguyen, T., Kerndt, C. C., & Dhamoon, A. S. (2023). Physiology, Blood Pressure Age Related Changes. In StatPearls. StatPearls Publishing.

Code Appendix:

```
knitr::opts_chunk$set(echo = FALSE)

library(tidyverse)
library(gtsummary)
library(knitr)
library(kableExtra)
library(mice)
library(gridExtra)

source("functions.R")
# read in data
framingham_df <- read.csv("framingham.csv")
nhanes_df <- read.csv("nhanes.csv") %>%
  mutate(SYSBP_UT = ifelse(BPMEDS == 0, SYSBP, 0),
         SYSBP_T = ifelse(BPMEDS == 1, SYSBP, 0))

framingham_df$S <- 1
nhanes_df$S <- 0


combined_df <- bind_rows(framingham_df, nhanes_df) %>%
  dplyr::select(-c(INCOME, RACE))

men_df <- combined_df %>% filter(SEX == "Male")
women_df <- combined_df %>% filter(SEX == "Female")

# Table with prevalence of missing NHANES data
data.frame(variable = names(colMeans(is.na(nhanes_df))),
           perc_missing = paste0(round(100*colMeans(is.na(nhanes_df)),1),"%"),
           row.names = c("Systolic Blood Pressure", "DBP", "Sex", "Age",
                         "Income", "Race", "BMI", "HDL Cholesterol", "Smoking",
                         "BP Medication",  "Total Cholesterol", "Diabetes",
                         "SBP if not treated", "SBP if treated", "S"))  %>%
  filter(variable %in% c("SEX", "AGE", "HDLC", "TOTCHOL", "SYSBP", "SYSBP_UT",
                         "SYSBP_T",  "BPMEDS",  "CURSMOKE", "DIABETES")) %>%
  slice(match(c("SEX", "AGE", "CURSMOKE", "DIABETES",
                "BPMEDS", "HDLC", "TOTCHOL", "SYSBP"), variable)) %>%
  dplyr::select(-variable) %>%
  kableExtra::kbl(format = "latex", caption = "Prevalence of Missing Data in NHANES Cohort",
                  col.names = "Percent Missing",
                  booktabs = T, linesep = "") %>%
    kableExtra::kable_styling(latex_options = c("HOLD_position"), font_size = 8)


nhanes_imp <- mice(nhanes_df, m = 5, seed = 1, print = FALSE)

nhanes_imp_ls <- vector("list", 5)
combined_imp_ls <- vector("list", 5)
for (i in 1:5) {

  nhanes_imp_ls[[i]] <- complete(nhanes_imp, i) %>%
```

```
    mutate(SYSBP_UT = ifelse(BPMEDS == 0, SYSBP, 0),
           SYSBP_T = ifelse(BPMEDS == 1, SYSBP, 0))

  combined_imp_ls[[i]] <- bind_rows(framingham_df, complete(nhanes_imp, i)) %>%
    dplyr::select(-c(SYSBP_UT, SYSBP_T, INCOME, RACE))

}


# Filter to each sex
framingham_df_men <- framingham_df %>% filter(SEX == "Male")
framingham_df_women <- framingham_df %>% filter(SEX == "Female")


### Fit Framingham Models ###

# 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= framingham_df_men, family= "binomial")


mod_women <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                   log(SYSBP_T+1)+CURSMOKE+DIABETES,
                 data= framingham_df_women, family= "binomial")


men_df_summary <- men_df %>%
  mutate(COHORT = ifelse(S == "1", "FHS", "NHANES"),
         SYSBP_UT = ifelse(SYSBP_UT == 0, NA, SYSBP_UT),
         SYSBP_T = ifelse(SYSBP_T == 0, NA, SYSBP_T))

tbl_summary(men_df_summary,
            by = COHORT,
            include = c(CVD, AGE, SYSBP, SYSBP_UT, SYSBP_T, DIABP, HDLC, TOTCHOL,
                        BMI, CURSMOKE, DIABETES, BPMEDS),
            label = c(TOTCHOL ~ "Total Cholesterol (mm/dL)",
                      AGE ~ "Age",
                      SYSBP ~ "Systolic Blood Pressure (mmHg)",
                      SYSBP_UT ~ "-- SBP if Not Treated",
                      SYSBP_T ~ "-- SBP if Treated",
                      DIABP ~ "Diastolic Blood Pressure (mmHg)",
                      CURSMOKE ~ "Cigarette Smoker",
                      DIABETES ~ "Diabetic",
                      BPMEDS ~ "Uses Anti-hypertensive Medication",
                      HDLC ~ "HDL Cholesterol (mg/dL)"),
            statistic = list(all_continuous() ~
                               c("{mean} ({sd})"),
                             all_categorical() ~
                               c("{p}%")),
            missing = "no") %>%
  modify_table_body(~.x %>%
                      mutate(
```

```
                        across(all_stat_cols(), ~gsub("NA%", "--
", .))
  )) %>%
   as_kable_extra(booktabs = TRUE,
                  caption = "Characteristics of Men in the FHS and NHANES Cohorts",
                 longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))


women_df_summary <- women_df %>%
  mutate(COHORT = ifelse(S == "1", "FHS", "NHANES"),
         SYSBP_UT = ifelse(SYSBP_UT == 0, NA, SYSBP_UT),
         SYSBP_T = ifelse(SYSBP_T == 0, NA, SYSBP_T))

tbl_summary(women_df_summary,
            by = COHORT,
            include = c(CVD, AGE, SYSBP, SYSBP_UT, SYSBP_T, DIABP, HDLC, TOTCHOL,
                        BMI, CURSMOKE, DIABETES, BPMEDS),
            label = c(TOTCHOL ~ "Total Cholesterol (mm/dL)",
                      AGE ~ "Age",
                      SYSBP ~ "Systolic Blood Pressure (mmHg)",
                      SYSBP_UT ~ "-- SBP if Not Treated",
                      SYSBP_T ~ "-- SBP if Treated",
                      DIABP ~ "Diastolic Blood Pressure (mmHg)",
                      CURSMOKE ~ "Cigarette Smoker",
                      DIABETES ~ "Diabetic",
                      BPMEDS ~ "Uses Anti-hypertensive Medication",
                      HDLC ~ "HDL Cholesterol (mg/dL)"),
            statistic = list(all_continuous() ~
                               c("{mean} ({sd})"),
                             all_categorical() ~
                               c("{p}%")),
            missing = "no") %>%
  modify_table_body(~.x %>%
                      mutate(
                        across(all_stat_cols(), ~gsub("NA%", "--
", .))
  )) %>%
   as_kable_extra(booktabs = TRUE,
                  caption = "Characteristics of Women in the FHS and NHANES Cohorts",
                 longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))
coef_table_men <- summary(mod_men)$coefficients %>%
  data.frame(row.names = c("Intercept",
                           "Log of HDL cholesterol",
                           "Log of total cholesterol",
                           "Log of age",
                           "Log of SBP if not treated",
                           "Log of SBP if treated",
                           "Smoking",
                           "Diabetes")) %>%
```

```
  dplyr::select(-z.value) %>%
  mutate(HR = exp(Estimate),
         LB = exp(Estimate - 1.96*`Std..Error`),
         UB = exp(Estimate + 1.96*`Std..Error`),
         CI = paste0("(", as.character(round(LB, 2)), " - ",
                     as.character(round(UB, 2)), ")")) %>%
  dplyr::select(Estimate, `Pr...z..`, HR, CI) %>%
  mutate(Estimate = round(Estimate, 2),
         `Pr...z..` = ifelse(`Pr...z..` < 0.001, "<0.001",
                             as.character(round(`Pr...z..`, 3))),
         HR = round(HR, 2))


coef_table_men[1,2:4] <- ""

coef_table_men %>%
  kableExtra::kbl(format = "latex",
                  caption = "Coefficients and Odds Ratios for Male Risk Model",
                  col.names = c("Estimate", "p-value", "OR", "95% CI"),
                  booktabs = T, linesep = "") %>%
    kableExtra::kable_styling(latex_options = c("HOLD_position"), font_size = 8)


coef_table_women <- summary(mod_women)$coefficients %>%
  data.frame(row.names = c("Intercept",
                           "Log of HDL cholesterol",
                           "Log of total cholesterol",
                           "Log of age",
                           "Log of SBP if not treated",
                           "Log of SBP if treated",
                           "Smoking",
                           "Diabetes")) %>%
  dplyr::select(-z.value) %>%
  mutate(HR = exp(Estimate),
         LB = exp(Estimate - 1.96*`Std..Error`),
         UB = exp(Estimate + 1.96*`Std..Error`),
         CI = paste0("(", as.character(round(LB, 2)), " - ",
                     as.character(round(UB, 2)), ")")) %>%
  dplyr::select(Estimate, `Pr...z..`, HR, CI) %>%
  mutate(Estimate = round(Estimate, 2),
         `Pr...z..` = ifelse(`Pr...z..` < 0.001, "<0.001",
                             as.character(round(`Pr...z..`, 3))),
         HR = round(HR, 2))


coef_table_women[1,2:4] <- ""

coef_table_women %>%
  kableExtra::kbl(format = "latex",
                  caption = "Coefficients and Odds Ratios for Female Risk Model",
                  col.names = c("Estimate", "p-value", "OR", "95% CI"),
                  booktabs = T, linesep = "") %>%
    kableExtra::kable_styling(latex_options = c("HOLD_position"), font_size = 8)
```

```r
predicted_men <- predict(mod_men, type = "response")
observed_men <- framingham_df_men$CVD
predicted_women <- predict(mod_women, type = "response")
observed_women <- framingham_df_women$CVD

# FHS brier score
brier_source_men <- mean((predicted_men - observed_men)^2)

brier_source_women <- mean((predicted_women - observed_women)^2)

auc_source_men <- estimate_auc_noweights(observed_men, predicted_men)

auc_source_women <- estimate_auc_noweights(observed_women, predicted_women)



brier_men <- rep(NA, 5)
brier_women <- rep(NA, 5)
auc_men <- rep(NA, 5)
auc_women <- rep(NA, 5)
for (i in 1:5) {

  # Get index values for each group
  index_M1 <- which(combined_imp_ls[[i]]$S == 1 & combined_imp_ls[[i]]$SEX == "Male")
  index_M0 <- which(combined_imp_ls[[i]]$S == 0 & combined_imp_ls[[i]]$SEX == "Male")
  index_F1 <- which(combined_imp_ls[[i]]$S == 1 & combined_imp_ls[[i]]$SEX == "Female")
  index_F0 <- which(combined_imp_ls[[i]]$S == 0 & combined_imp_ls[[i]]$SEX == "Female")

  # Get weights
  weights = get_weights(combined_imp_ls[[i]])

  # Calculate brier score for men
  brier_men[i] <- estimate_brier(model = mod_men,
                                 combined_df = filter(combined_imp_ls[[i]],
                                                      SEX == "Male"),
                                 observed = observed_men,
                                 predicted = predicted_men,
                                 n_target = length(index_M0))

  auc_men[i] <- estimate_auc(combined_df = filter(combined_imp_ls[[i]],
                                                  SEX == "Male"),
                             observed = observed_men,
                             predicted = predicted_men)

  # Calculate brier score for women
  brier_women[i] <- estimate_brier(model = mod_women,
                                   combined_df = filter(combined_imp_ls[[i]],
                                                        SEX == "Female"),
                                   observed = observed_women,
                                   predicted = predicted_women,
                                   n_target = length(index_F0))

  auc_women[i] <- estimate_auc(combined_df = filter(combined_imp_ls[[i]],
```

```
                                                SEX == "Female"),
                              observed = observed_women,
                              predicted = predicted_women)

}

brier_male_estimand = mean(brier_men)
brier_female_estimand = mean(brier_women)
auc_male_estimand = mean(auc_men)
auc_female_estimand = mean(auc_women)
sim_results <- read.csv("simulation_results.csv")
n_sim = nrow(sim_results)/100


# Brier Score

avg_results <- sim_results %>%
  group_by(rho, gamma) %>%
  reframe(avg_brier_male = mean(brier_male),
          avg_brier_female = mean(brier_female),
          avg_auc_male = mean(auc_male),
          avg_auc_female = mean(auc_female))

performance_brier_rho <- sim_results %>%
  group_by(rho) %>%
  reframe(bias_male = relative_bias(brier_male, brier_male_estimand),
          bias_se_male = relative_bias_se(n_sim, brier_male, brier_male_estimand),
          bias_female = relative_bias(brier_female, brier_female_estimand),
          bias_se_female = relative_bias_se(n_sim, brier_female, brier_female_estimand))

performance_brier_gamma <- sim_results %>%
  group_by(gamma) %>%
  reframe(bias_male = relative_bias(brier_male, brier_male_estimand),
          bias_se_male = relative_bias_se(n_sim, brier_male, brier_male_estimand),
          bias_female = relative_bias(brier_female, brier_female_estimand),
          bias_se_female = relative_bias_se(n_sim, brier_female, brier_female_estimand))


brier_rho_plot <- ggplot(performance_brier_rho) +
  geom_line(aes(x = rho, y = bias_male)) +
  geom_errorbar(aes(x = rho, ymax = bias_male + bias_se_male,
                    ymin = bias_male - bias_se_male), alpha = 0.5, width = 0.01) +
  geom_line(aes(x = rho, y = bias_female), linetype = "dashed") +
  geom_errorbar(aes(x = rho, ymax = bias_female + bias_se_female,
                    ymin = bias_female - bias_se_female), alpha = 0.5, width = 0.01) +
  scale_x_continuous(breaks = seq(0, 0.9, 0.1)) +
  labs(x = expression(rho), y = "Relative Bias") +
  lims(y = c(0.25, 1.75)) +
  theme_bw() +
  theme(panel.grid.minor = element_blank())

brier_gamma_plot <- ggplot(performance_brier_gamma) +
  geom_line(aes(x = gamma, y = bias_male)) +
```

```
  geom_errorbar(aes(x = gamma, ymax = bias_male + bias_se_male,
                    ymin = bias_male - bias_se_male), width = 0.01,
                alpha = 0.5) +
  geom_line(aes(x = gamma, y = bias_female), linetype = "dashed") +
  geom_errorbar(aes(x = gamma, ymax = bias_female + bias_se_female,
                    ymin = bias_female - bias_se_female, width = 0.01),
                alpha = 0.5) +
  scale_x_continuous(breaks = seq(0, 0.9, 0.1)) +
  labs(x = expression(gamma), y = "") +
  lims(y = c(0.25, 1.75)) +
  theme_bw() +
  theme(panel.grid.minor = element_blank())


grid.arrange(brier_rho_plot, brier_gamma_plot, ncol = 2,
             top = "Figure 1. Brier Score Bias for Men (solid) and Women (dashed)")

# AUC
performance_auc_rho <- sim_results %>%
  group_by(rho) %>%
  reframe(bias_male = relative_bias(auc_male, auc_male_estimand),
          bias_se_male = relative_bias_se(n_sim, auc_male, auc_male_estimand),
          bias_female = relative_bias(auc_female, auc_female_estimand),
          bias_se_female = relative_bias_se(n_sim, auc_female, auc_female_estimand))

performance_auc_gamma <- sim_results %>%
  group_by(gamma) %>%
  reframe(bias_male = relative_bias(auc_male, auc_male_estimand),
          bias_se_male = relative_bias_se(n_sim, auc_male, auc_male_estimand),
          bias_female = relative_bias(auc_female, auc_female_estimand),
          bias_se_female = relative_bias_se(n_sim, auc_female, auc_female_estimand))

auc_rho_plot <- ggplot(performance_auc_rho) +
  geom_line(aes(x = rho, y = bias_male)) +
  geom_errorbar(aes(x = rho, ymax = bias_male + bias_se_male,
                    ymin = bias_male - bias_se_male), alpha = 0.5, width = 0.01) +
  geom_line(aes(x = rho, y = bias_female), linetype = "dashed") +
  geom_errorbar(aes(x = rho, ymax = bias_female + bias_se_female,
                    ymin = bias_female - bias_se_female), alpha = 0.5, width = 0.01) +
  scale_x_continuous(breaks = seq(0, 0.9, 0.1)) +
  labs(x = expression(rho), y = "Relative Bias") +
  lims(y = c(-0.05, 0.3)) +
  theme_bw() +
  theme(panel.grid.minor = element_blank())

auc_gamma_plot <- ggplot(performance_auc_gamma) +
  geom_line(aes(x = gamma, y = bias_male)) +
  geom_errorbar(aes(x = gamma, ymax = bias_male + bias_se_male,
                    ymin = bias_male - bias_se_male), width = 0.01,
                alpha = 0.5) +
  geom_line(aes(x = gamma, y = bias_female), linetype = "dashed") +
  geom_errorbar(aes(x = gamma, ymax = bias_female + bias_se_female,
                    ymin = bias_female - bias_se_female, width = 0.01),
```

```
                alpha = 0.5) +
  scale_x_continuous(breaks = seq(0, 0.9, 0.1)) +
  labs(x = expression(gamma), y = "") +
  lims(y = c(-0.05, 0.3)) +
  theme_bw() +
  theme(panel.grid.minor = element_blank())

grid.arrange(auc_rho_plot, auc_gamma_plot, ncol = 2,
             top = "Figure 2. AUC Bias for Men (solid) and Women (dashed)")

data.frame(row.names = c("Males\n", "Females\n", "Males", "Females"),
           theta = c(0.118, 0.064, 0.758, 0.824),
           theta_zero = c(0.156, 0.096, 0.773, 0.790),
           theta_opt = c("0.149 ($\\rho = 0.3$, $\\gamma=0.5$)",
                         "0.086 ($\\rho = 0.2$, $\\gamma=0.9$)",
                         "0.762 ($\\rho = 0.0$, $\\gamma=0.9$)",
                         "0.824 ($\\rho = 0.7$, $\\gamma=0.1$)")) %>%
  kableExtra::kbl(format = "latex",
                  col.names = c("$\\theta$",
                                "$\\hat\\theta_{\\text{none}}$",
                                "$\\hat\\theta_{\\text{opt}}$"),
                  escape = FALSE, booktabs = T,
                  caption = "Comparison of Association Assumptions") %>%
  group_rows(group_label = "Brier Score", start_row = 1, end_row = 2) %>%
  group_rows(group_label = "AUC", start_row = 3, end_row = 4) %>%
    kableExtra::kable_styling(latex_options = c("HOLD_position"), font_size = 10)
```