Project 3: Transportability Analysis

Github Link

Zihan Zhou

2023-12-10

Abstract

Aim: This study aims to assess the transportability of a gender-specific risk score model for cardiovascular diseases (CVD), built using data from the Framingham Heart Study (FHS), to the population underlying the NHANES (National Health and Nutrition Examination Survey) survey data through a simulation study.

Method: The Brier Score is used to measure predictive accuracy for logistic prediction models for men and women based on Framingham data. These models were then applied to both the NHANES population and a simulated NHANES population based on summary statistics, and assessed by Brier score with adjustment for weights.

Result: The average Brier scores for FHS are 0.1993 for men and 0.1184 for women. When the model is applied to NHANES, the Brier Scores change to 0.2013 for men and 0.1374 for women. The Brier scores for the simulated NHANES data have a mean of 0.2123 with a standard deviation of 0.0263 for men, and a mean of 0.1278 with a standard deviation of 0.0149 for women.

Conclusions: The analysis reveals that the model for women achieves a higher Brier Score than the model for men across all three datasets, suggesting that the transportability of the model is better for women than for men. They highlight the challenges in transporting prediction models across diverse populations, emphasizing the impact of varying covariate distributions on model performance. The transportability of the Brier Score is constrained by two identification conditions. Future research could focus on additional performance metrics, such as the AUC, and on developing methods to better address missing data.

1. Introduction

Assessing the performance of prediction models is one of the crucial aspects when developing a model. Usually, the development of prediction models is driven by predictions within a specific target population. In the healthcare system, these prediction models are desired to be deployed to identify individuals at high risk for certain diseases across different populations. The data used to develop the model, referred to as the source study data, often come from randomized trials [1], large observational databases [2], or prospective cohort studies. However, these sources may not represent random samples of the target population, leading to potential disparities in data distribution. This discrepancy makes it difficult to derive the accurate assessment of a model's performance on the target population, as performance metrics derived from the source population may not translate directly to the target population. For instance, the Framingham ATP-III model [3], designed to predict the 10-year risk of cardiovascular events and primarily developed from data with predominantly white participants, has shown limited generalizability to multi-ethnic populations.

When assessing the performance on external dataset, challenges may arise when only covariate data are available. Due to the differences in data distribution compared to the source data, the outcome data can also have different distribution. For example, the Framingham Heart study [4] provides both covariate and outcome data, while the National Health and Nutrition Examination Survey (NHANES) [5] only has covariate data. This lack of outcome information from the target population limits the ability to develop or assess prediction models using data solely from the target population. In recent years, several methods have been introduced to evaluate prediction model performance in a target population, or to transfer performance measures from the source to the target population.

This project will apply the Brier score as a measure of transportability, a newly developed method that uses covariate and outcome data from the source population to bridge the differences in data distributions between the two populations [6]. The approach will be applied to two gender-based risk score models for cardiovascular disease (CVD) developed using Framingham Heart Study data. The objective of the study is to estimate the models' performance in a new target population represented by the NHANES survey data. When only summary statistics of the target population are available, simulation studies will be utilized, simulating the population based on the target population's statistics.

2. Data

2.1 Framingham study

The Framingham Heart Study [7], conducted in Framingham, Massachusetts, is a long-term prospective investigation into the causes of cardiovascular disease (CVD) among a population of free living subjects. This study in epidemiology was pioneering as it was the first study to prospectively examine cardiovascular disease, introducing the concept of risk factors and their joint effects. The study started in 1948 and a total of 5,209 subjects were initially enrolled in the study. All research participants were constantly monitored for the emergence of CVD events and mortality. CVD is defined by the Framingham study as a composite of CHD (coronary death, myocardial infarction, coronary insufficiency, and angina), cerebrovascular events (including ischemic stroke, hemorrhagic stoke, and transient ischemic attack), peripheral artery disease (intermittent claudication), and heart failure.

Eligible participants included those who attended the 11th biennial check-up of the original cohort between 1968 and 1971 (a period when high-density lipoprotein [HDL] cholesterol measurements were available), or the first (1971-1975) or third (1984-1987) check-ups of the Offspring cohort, were aged between 30 and 74, and did not have CVD [4].

2.2 NHANES

The National Health and Nutrition Examination Survey (NHANES) [5] is a significant U.S. program aimed at assessing the health and nutritional status of adults and children. Managed by the National Center for Health Statistics (NCHS) under the CDC, NHANES combines interviews with physical examinations. Initiated in the early 1960s, the program has become a continuous effort since 1999, focusing on various health and nutrition topics. It examines about 5,000 individuals each year from a nationally representative sample, covering 15 counties each year.

During the home interview, participants answer questions related to their health condition, medical history, and eating patterns, while in the health examination stage, they go thorough medical and dental assessments, accurate physiological measurements, and a range of detailed laboratory examinations, all conducted by trained medical staff. This comprehensive strategy guarantees an in-depth insight into the health profiles of the participants, building a strong foundation for thorough health and nutritional studies.

2.3 Summary Statistics

The Framingham and NHANES datasets have been cleaned and transformed to include the same covariates, with the eligibility criteria, particularly the age limitation between 30 to 74 from the Framingham study, applied to the NHANES sample. Both datasets are pre-processed to select covariates commonly used in CVD prediction models.

Table 1 presents a summary of the variables used in this transportability analysis, where eligibility criteria have been applied to the NHANES data, and no other transformations or imputations have been performed. In the Framingham source population, cardiovascular disease (CVD) events were recorded for 33% of men (360) and 17% of women (242); however, this CVD data is not available in the NHANES dataset. This table also highlights the differences in the distribution of risk factors by gender and between the two populations, based on the p-value calculated by Pearson's Chi-squared test and Wilcoxon rank sum test. Compared to NHANES participants (188 mg/dL for men and 196 mg/dL for women), Framingham participants have higher average total cholesterol levels (226 mg/dL for men and 246 mg/dL for women). Framingham study participants also showed a higher rate of current cigarette smoking among both men (39%) and women (31%) compared to NHANES participants (24% for men and 16% for women). Differences in diabetic status and the use of anti-hypertensive medication are observed, with the NHANES population showing higher rates in both categories, but only the difference in diabetic status was statistically significant between genders. The differences in these risk factors by gender shows the necessity to model based on gender while the differences in covariates highlights the challenge when transporting the risk score model from Framingham population to NHANES population.

Table 1: Summary of the variables used in the transportability analysis of CVD predictio model

	Framingham			NHANES		
SEX	**1**, N = 1,094	**2**, N = 1,445	**p-value**	**1**, N = 1,961	**2**, N = 2,099	**p-value**
CVD	360 (33%)	242 (17%)	<0.001			
Serum Total Cholesterol (mg/dL)	226.44 (41.49)	246.32 (45.51)	<0.001	188.40 (41.68)	195.64 (40.13)	<0.001
AGE	60.01 (8.18)	60.55 (8.40)	0.13	52.92 (12.71)	51.91 (12.56)	0.009
Systolic Blood Pressure	138.94 (20.89)	139.94 (23.71)	0.6	128.29 (17.67)	125.78 (20.24)	<0.001
Current Cigarette Smoking	425 (39%)	445 (31%)	<0.001	474 (24%)	337 (16%)	<0.001
Diabetic	96 (8.8%)	95 (6.6%)	0.037	358 (18%)	317 (15%)	0.007
Use of Anti-hypertensive Medication	123 (11%)	259 (18%)	<0.001	604 (33%)	629 (32%)	0.4
High Density Lipoprotein Cholesterol (mg/dL)	43.63 (13.37)	53.07 (15.67)	<0.001	47.83 (14.07)	57.60 (15.84)	<0.001
BMI	26.25 (3.47)	25.55 (4.22)	<0.001	29.90 (6.56)	30.82 (8.19)	0.040
Systolic Blood Pressure (No BPMEDS)	121.04 (46.69)	111.49 (55.89)	<0.001	79.19 (61.47)	77.19 (59.19)	<0.001
Systolic Blood Pressure (On BPMEDS)	17.90 (50.93)	28.45 (61.53)	<0.001	39.34 (61.49)	37.61 (61.67)	0.4

[1] n (%); Mean (SD)

[2] Pearson's Chi-squared test; Wilcoxon rank sum test

[3] Mean (SD); n (%)

[4] Wilcoxon rank sum test; Pearson's Chi-squared test

2.3 Missing Data

There is a lot of missing data in both datasets. When developing the risk score model based on the Framingham study, only complete records are used. As a result, the challenges of missing data in the Framingham study will not be considered. When transferring the model to NHANES, multiple imputation will be employed. Table 2 presents the percentage of missing data in NHANES after applying the eligibility. Some variables seem to exhibit a similar pattern of missing data. The `HDLC` and `TOTCHOL` variables have identical missing percentages of 10.52%.

Table 2: Summary of Missing Values for NHANES

Variable	Number	Pct
SYSBP_UT	646	15.91%
SYSBP	645	15.89%
SYSBP_T	448	11.03%
HDLC	427	10.52%
TOTCHOL	427	10.52%
BPMEDS	248	6.11%
BMI	238	5.86%
DIABETES	1	0.02%

3. Transportability Analysis Methods

3.1. Multiple Imputation

Multiple imputation will be used to address the missing data in the NHANES dataset. This method creates several imputed datasets, thus introducing variability in the imputed values to reflect the uncertainty around the true values. The imputation will apply algorithms repetitively to generate values for the missing data. Then each dataset undergoes standard statistical analysis since it is complete. In this project, five complete NHANES datasets will be produced for analysis, the pooled Brier Score will be used.

3.2 Models

In this project, two gender-based models developed using the Framingham Dataset for predicting cardiovascular disease (CVD) will be adapted for use with the NHANES population. The NHANES data lacks relevant CVD outcomes with only covariates data. The models will be evaluated on three different datasets: the original Framingham Dataset, the NHANES dataset, and a simulated NHANES dataset. This evaluation aims to assess the transportability of the CVD risk score model. Since CVD is a binary outcome in the model, logistic regression models are used.

For Men:

$$\begin{aligned}\operatorname{logit}(P(\text{CVD})) = {} & \beta_0 + \beta_1 \log(\text{HDLC}) + \beta_2 \log(\text{TOTCHOL}) \\ & + \beta_3 \log(\text{AGE}) + \beta_4 \log(\text{SYSBP\_UT} + 1) \\ & + \beta_5 \log(\text{SYSBP\_T} + 1) + \beta_6 \text{CURSMOKE} + \beta_7 \text{DIABETES}\end{aligned}$$

For Women:

$$\begin{aligned}\operatorname{logit}(P(\text{CVD})) = {} & \gamma_0 + \gamma_1 \log(\text{HDLC}) + \gamma_2 \log(\text{TOTCHOL}) \\ & + \gamma_3 \log(\text{AGE}) + \gamma_4 \log(\text{SYSBP\_UT} + 1) \\ & + \gamma_5 \log(\text{SYSBP\_T} + 1) + \gamma_6 \text{CURSMOKE} + \gamma_7 \text{DIABETES}\end{aligned}$$

Here, $\operatorname{logit}(P(\text{CVD}))$ represents the log-odds of cardiovascular disease (CVD). HDLC stands for High-Density Lipoprotein Cholesterol (mg/dL), TOTCHOL for Serum Total Cholesterol (mg/dL), and SYSBP for Systolic Blood Pressure, adjusted for the use of anti-hypertensive medication (BOMEDS). CURSMOKE indicates current cigarette smoking status (0 = Not a current smoker, 1 = Current smoker), and DIABETES represents the presence of diabetes based on the criteria of the first exam treated, or a casual glucose level of 200 mg/dL or more (0 = Not diabetic, 1 = Diabetic). The coefficients β_i and γ_i are for the men's and women's models respectively. The data will be split into 70-30 train-test sets. The model will be trained on the training data and then evaluated on the test data.

3.3 Simulation

The simulation study is presented in the ADEMP structure [8].

Aim

The aim of this simulation study is to assess the transportability of the cardiovascular disease (CVD) event prediction model, originally developed from the Framingham study data, to the NHANES national population. This assessment will be conducted using the Brier score as a measure of predictive accuracy when only summary data from the NHANES are available.

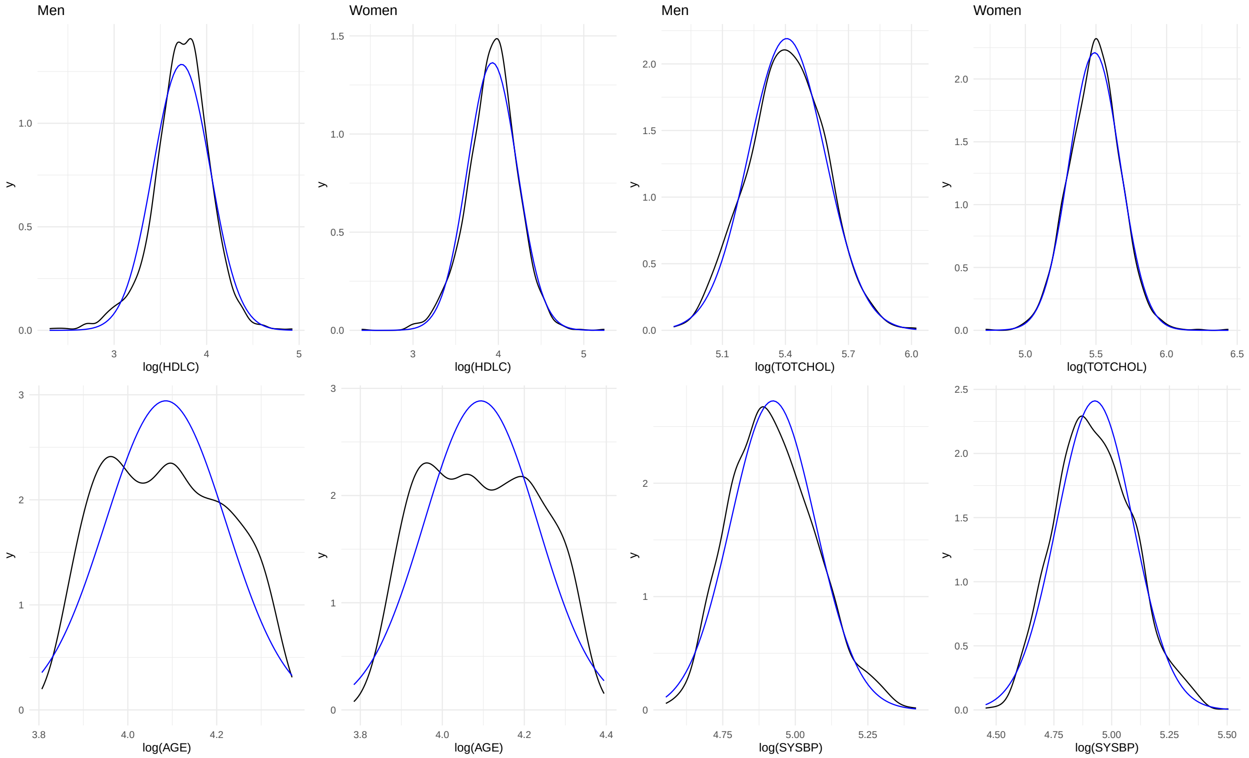
Data-Generating Mechanism

When simulating the data, we assume that individual level data is not available from the target population NHANES and only summary statistics in Table 1 are available. When adjusting for the distribution of NHANES covariates, covariate data from the source population, Framingham, are used. The Shapiro-Wilk normality test is utilized to determine if simulations using a normal distribution are applicable. The results are consistently much smaller than the significance level of 0.05, leading to the rejection of the normality assumption. Therefore, we consider alternative distributions.

After exploratory analysis, it can be observed that the distribution are right-skewed. In the following plots, the black lines represent the densities of continuous variables of Framingham and the overlay blue lines are for log normal distribution using summary statistics with mean and standard deviation. The plots show that the distributions of the actual data and the simulated data are very close, except for age, which shows a small deviation. The data-generation is done separately for male and female. As a result, log-normal distribution is then used to generate continuous variables in the simulated NHANES dataset based on the summary statistics.

For binary variables, they are also associated to continuous variables, so random samples will be drawn from a multivariate normal distribution for all continuous and binary variables. Then based on the summary statistics for binary variables, the generated continuous value will be converted into binary value based on their quantile. Each simulated dataset contains 1,961 men and 2,099 women. The size of the simulation is $n_{sim} = 1000$.

Figure 1: Densities of actual Framingham study and simulated using Log−normal distribution



Estimands

The estimands used in this simulation study is the Brier Scores in the source population and target population (which is equivalent to the MSE for binary outcomes). In the source population, the Brier Score can be calculated using the mean squared error (MSE) $(Y - g(X))^2$, which quantifies the discrepancy between the observed outcome Y and the model Pr[Y=1|X, D=0] derived prediction $g(X)$, where X is the covaraites.

In target population, the outcome variable is missing, which presents a challenge for calculations. Let S be an indicator for the population from which data are obtained, with S = 1 representing the source population Framingham and S = 0 representing the target population NHANES. The term $n = n_{source} + n_{target}$ is used to denote the total number of observations in the composite dataset, which consists of the data from both the source and target poopulation. Then the dataset is divided into a training set and a test set. Let D be an indicator distinguishing between the training and test data within the population. Therefore, the Brier Score for the target population is [6]:

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

where $\hat{o}(X)$ is an estimator for the inverse-odds weights in the test set,

$$\frac{\Pr[S = 0|X, D_{test=1}]}{\Pr[S = 1|X, D_{test=1}]}.$$

The weights $\hat{o}(X)$ can be obtained by fitting a logistic regression model for the probability of the participants from the source population conditional on covariates $\Pr[S = 1|X, D_{test=1}]$.

Methods

Two logistic regression models presented in section 3.2 for male and female are fitted to predict CVD. Continuous variables are log-transformed. These models are developed on the training dataset and then evaluated on the test dataset.

Performance Measures

The transportable ability is assessed by comparing the estimated Brier scores between the source population Framingham study, the NHANES population and the simulated NHANES population.

4. Results

Table 3 summarizes the Brier Scores for the cardiovascular disease (CVD) risk prediction model across three datasets: Framingham, NHANES, and Simulated NHANES. It is observed that for both genders, the Brier Scores and their standard deviations are consistently higher in men than in women.

A lower Brier Score suggests a better accuracy of the model. For the Framingham dataset, the Brier Scores were 0.1993 for men and 0.1148 for women, suggesting more precise predictions for women.

When transporting this model to the NHANES data, there was an increase in Brier Scores, indicating less accuracy compared to the original Framingham data. The average Brier Score for men across the five imputation dataset is 0.2013 while for women is was 0.1374. The model for women still performs better than men. The standard deviation of Brier scores for women is smaller than men, suggesting more stable predictions.

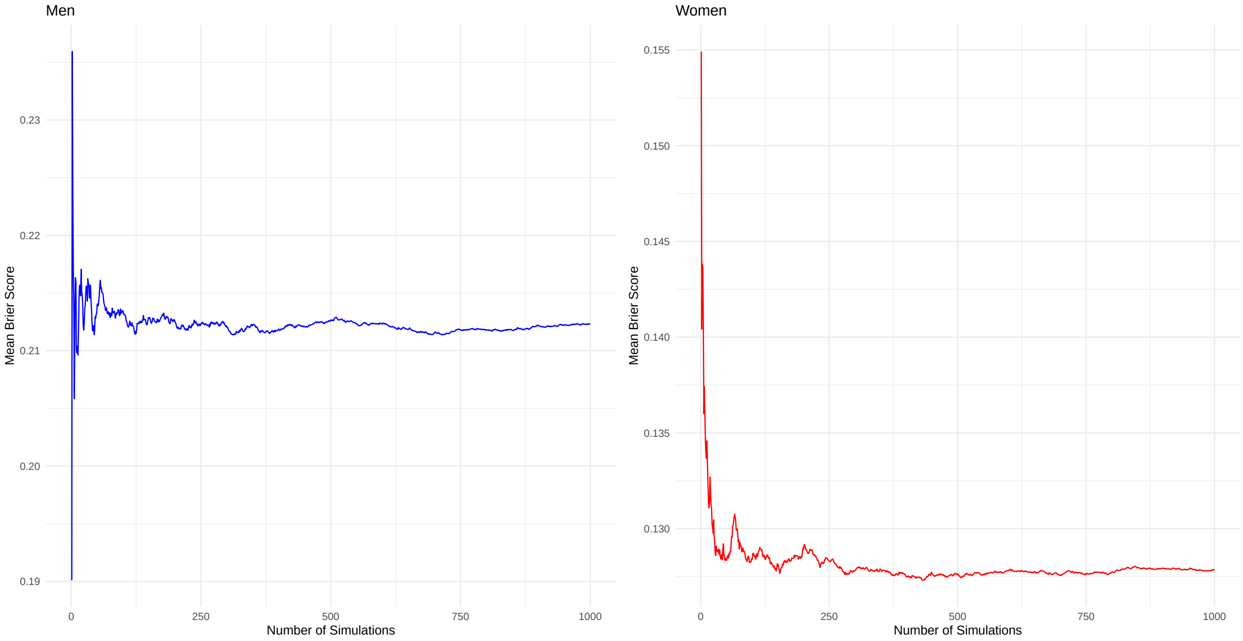
Simulations based on summary statistics yielded the smallest Brier Scores among the datasets, with the mean Brier Score for the 1000 simulated datasets being 0.2123 for men and 0.1278 for women, suggesting better performance of model for women on the simulated data. The standard deviation for women remained smaller than for men. These results indicate that the model's transportability for women is potentially better than for men in our studies.

Table 3: Summary of the Brier Scores

	Statistics	Framingham	NHANES	Simulated NHANES (n = 1000)
Men	Mean	0.1993	0.2013	0.2123
	SD	-	0.0224	0.0263
Women	Mean	0.1148	0.1374	0.1278
	SD	-	0.0083	0.0149

Figure 2 further illustrates the results of the simulation study, plotting the changes in Brier Scores against the number of simulations for both men and women. It shows the variability and convergence of the mean Brier Score as the number of simulations increases. The figure on the left for men, displays initial fluctuations. However, as the number of simulations is over 300, the mean Brier Score stabilizes and converges to a value around 0.2123. The figure on the right for women, also demonstrates initial variability, but the mean Brier Score stabilizes and converges more quickly than men, eventually reaching a value around 0.1278. It indicates a more consistent prediction model for women compared to men, aligning with our results in Table 3 regarding the standard deviation.

Figure 2: Changes of mean Brier Score with the Number of Simulation



5. Discussion

In this project, the Brier Score is used to estimate the transferability of CVD prediction models to a target population during model development when outcome and covariate data are available from the source population, and only covariate data are available from the target population. Two logistic models for men and women have been developed using population data from the Framingham study. The model performance is estimated without specifying the model in the NHANES population. The formula for transportability analysis takes into account both the source and target populations.

Our results indicate that across all three datasets, the model consistently performs better for women than for men, suggesting underlying differences in CVD risk-related covariates. The model for women demonstrates a higher capacity for application across various female populations.

When comparing the results across the datasets, it shows that the Brier Score of model for men on the simulated NHANES population are the highest, while for women on the NHANES population are the highest. The model performs better on Framingham data than on NHANES data because it is developed based on the former, and the covariates of the NHANES data are significantly different from those of the Framingham data. The model can represent the Framingham data more accurately than the NHANES data since the measures of model performance are essentially an average of the covariate distribution. Therefore, when the distributions of the covaraites are different, the predictions will deviate.

The simulated data performs best for women might be due to that the simulated data is not able capture the full complexity of the real-world data. In our simulations, only log-normal distributions are used. Although the density plots for most covariates are close, age shows a deviation, and the log-normal distribution has fewer outliers compared to real-world data. Moreover, since we assume we do not have the individual data of NHANES when simulating, we can only refer to the distribution from the Framingham study. However, the actual distribution in NHANES may not follow the covariate distributions observed in Framingham.

There are also some limitations to our project. Firstly, the implemented Brier Score for transportability relies on two identifiable conditions, A1 and A2 [6]: A1 needs the independence of the outcome Y and the population S given the covariates, and A2 is positivity, suggesting $\Pr[S = 1|X = x] > 0$ for every x where the joint density of $f(X = x, S = 0) \neq 0$. These conditions might not be satisfied in other datasets, which limits the usage of this metric. Future studies could consider additional performance measures, such as AUC

(Area Under the ROC Curve), for a more comprehensive assessment of the model. Moreover, alternative distributions and methods could be explored for simulation purposes. The issue of missing data in the Framingham population has not been considered in our projct, thus new transportability analysis tools could also be developed to address this problem.

References

[1] R. Pajouheshnia, R. H. H. Groenwold, L. M. Peelen, J. B. Reitsma, and K. G. M. Moons, “When and how to use data from randomised trials to develop or validate prognostic models,” *BMJ*, vol. 365, p. l2154, May 2019, doi: 10.1136/bmj.l2154.

[2] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, “Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review,” *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 198–208, Jan. 2017, doi: gffmc5.

[3] National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III), “Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report,” *Circulation*, vol. 106, no. 25, pp. 3143–3421, Dec. 2002.

[4] R. B. D’Agostino *et al.*, “General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study,” *Circulation*, vol. 117, no. 6, pp. 743–753, Feb. 2008, doi: dthnkd.

[5] “NHANES - National Health and Nutrition Examination Survey Homepage.” https://www.cdc.gov/nchs/nhanes/index.htm.

[6] J. A. Steingrimsson, C. Gatsonis, B. Li, and I. J. Dahabreh, “Transporting a Prediction Model for Use in a New Target Population,” *American Journal of Epidemiology*, vol. 192, no. 2, pp. 296–304, Feb. 2023, doi: gs6sv9.

[7] W. B. Kannel, P. A. Wolf, R. J. Garrison, L. A. Cupples, and R. B. D’Agostino, *The Framingham study: An epidemiological investigation of cardiovascular disease / Section 34 : Some risk factors related to the annual incidence of cardiovascular disease and death using pooled repeated biennial measurements : Framingham heart study, 30 year followup / L. Adrienne Cupples and Ralph B. D’Agostino.* Bethesda, Md.: National Heart, Lung and Blood Institute, 1987.

[8] T. P. Morris, I. R. White, and M. J. Crowther, “Using simulation studies to evaluate statistical methods,” *Statistics in Medicine*, vol. 38, no. 11, pp. 2074–2102, May 2019, doi: 10.1002/sim.8086.