Tranportability Case-study and Simulation Analysis

Yu Yan

2023-11-29

Abstract

Background: Risk prediction models like the Framingham Heart Study are integral to clinical decision-making yet face hurdles when applied across diverse populations. This case study employs transportability analysis, following the methodology by Dr. Steingrimsson et al., to adapt the Framingham model for the NHANES target population. It also aims to compare the effectiveness of data-based and simulation-based transportability approaches.

Methods: This dual-goal analysis commenced with the evaluation of the Framingham model using NHANES data, focusing on performance metrics such as the Brier Score. The first approach involved direct application and modification of the model using inverse-odds weighting on combined data sets. The second goal introduced a simulation-based approach, generating data reflective of individual-level statistics from summary data to test transportability. Monte Carlo simulation methodology was employed and simulation bias are reported for evaluation.

Results: The study observed high predictive accuracy within the NHANES cohort, with gender-based analysis showing higher accuracy for females. The simulation-based approach, involving varying data generation mechanisms, aimed to reproduce individual-level data from summary statistics, providing insights into model applicability and performance under different scenarios. The bias in estimates from the simulation approach was compared with the data-based approach to assess the efficacy and feasibility of each method.

Conclusions: Findings suggest that both transportability approaches can effectively adapt the Framingham model to the NHANES population with minimal bias, highlighting the potential of simulation-based methods in scenarios where individual data is unavailable. This comparative study underscores the importance of considering various data generation methods in transportability analysis, ultimately broadening the applicability of predictive models in healthcare.

Introduction

Transferring a health risk prediction model from one group of people to another can be challenging, especially when the groups are very different. The well-known Framingham Heart Study has created a model that predicts heart health risks, but it's mostly been used on people within the study. Our goal is to see if this model can also work well for people in the NHANES study, which collects health and nutrition data from a wide range of Americans but doesn't have longitudinal heart related health outcomes

This study will use comparisons between data-based and simulation based approaches to estimate how well the Framingham model can predict health outcomes for the NHANES group. We'll do this by using the detailed health information from NHANES, along with the patterns of heart health outcomes from the Framingham study, to create a simulated set of results. In doing so, we'll also carefully apply the Framingham study's criteria to the NHANES data. Our analysis aims to show how well the Framingham heart risk model can be adapted for use with different groups of people.

Table 1: Summary Statistics of Nhance(0) and Framinghan(1) data

	Population Membership		
Variable	**0**, N = 3,904	**1**, N = 2,348	**p-value**
TOTCHOL	5.24 (5.11, 5.38)	5.46 (5.33, 5.58)	<0.001
AGE	3.99 (3.74, 4.14)	4.08 (3.97, 4.17)	<0.001
SYSBP	4.84 (4.74, 4.91)	4.91 (4.80, 5.02)	<0.001
HDLC	3.91 (3.71, 4.11)	3.87 (3.66, 4.04)	<0.001
BMI	3.37 (3.23, 3.53)	3.24 (3.15, 3.33)	<0.001
BPMEDS			<0.001
0	2,471 (67%)	2,008 (86%)	
1	1,193 (33%)	340 (14%)	
DIABETES			<0.001
0	3,259 (83%)	2,178 (93%)	
1	644 (17%)	170 (7.2%)	
SEX			<0.001
0	1,889 (48%)	1,022 (44%)	
1	2,015 (52%)	1,326 (56%)	
CURSMOKE			<0.001
0	3,117 (80%)	1,498 (64%)	
1	787 (20%)	850 (36%)	

[1] Median (IQR); n (%)

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

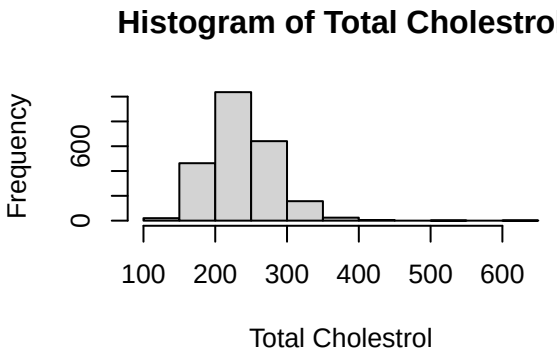
Data Processing

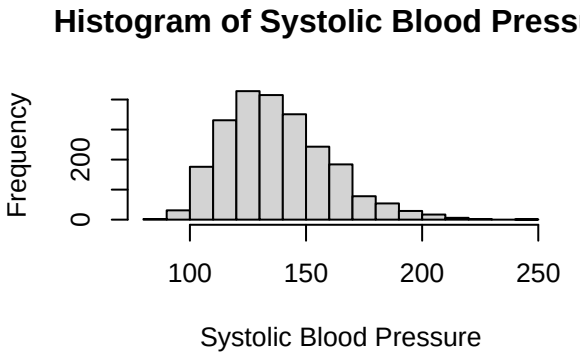
In the data processing part, since the Framingham data is provided as filtered and complete, we will create a new variable of source and give it a value of 1 indicate it as source population, and filter age range as 30-74, indicated by B. D'AgostinoSr et.al(D'Agostino et al. 2008). For the nhanes data, we will do some processing as follows: filter out observations whose age is above 30 and below 74 as the eligibility criteria that matches the setting of Framingham heart study, and then created both 'SYSBP_UT' and 'SYSBP_T' the same way in the processing of Framingham data. Lastly we added 'source' and give it a value of 0 indicating this is target data. As the model that we are evaluating is stratified by different sex, we will also divide both data by sex as subsets, and perform transportability analysis seperately.

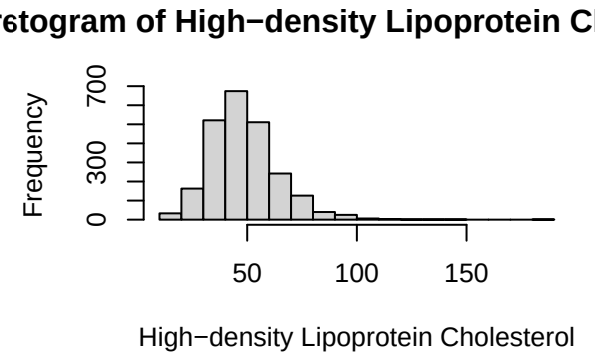
The followings are demographics of the two datasets after preliminary processing:

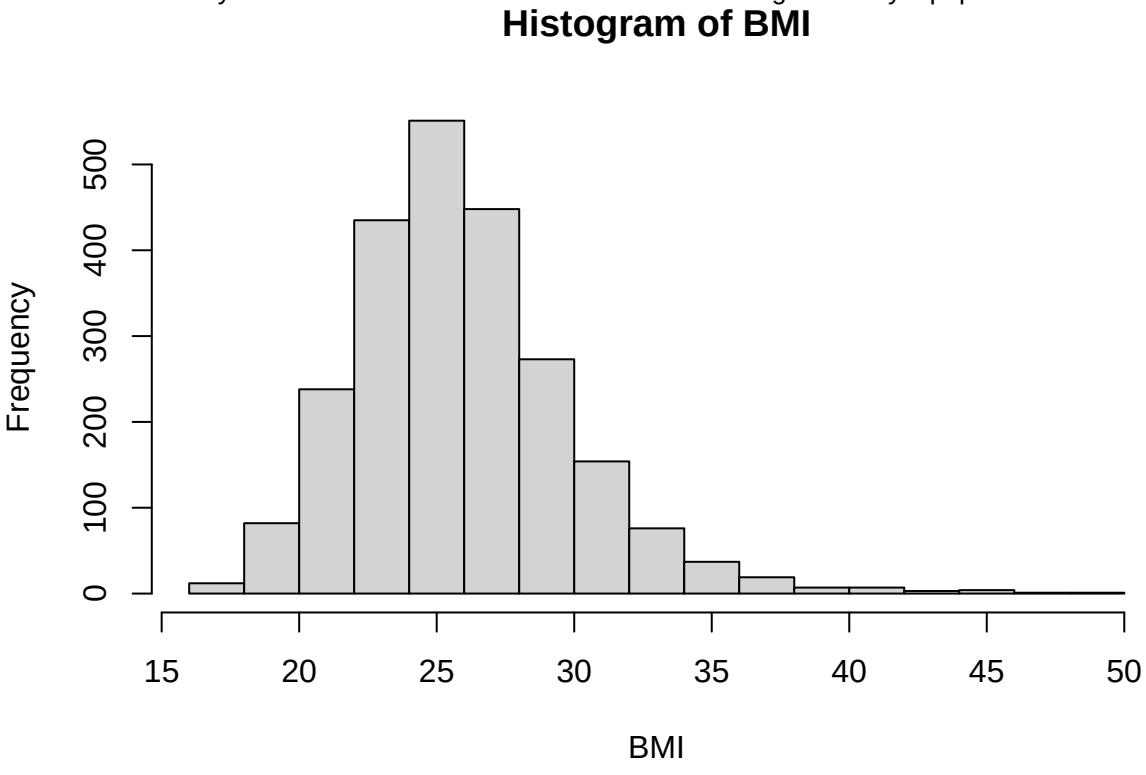
In the table of summary statistics of complete case Nhanes and Framinghan, We displayed several crucial variables that we filtered out. Those variables are basically the ones that are used by the CVD prediction model that we are evaluating. All of the differences between them seem statistically significant.

We also displayed histograms of continuous variables from Framingham dataset as a reference. One of the assumptions of the transportability analysis is related to the distribution of covarites in both target and source population. As later stage in the report, we will rely on such distributions to simulate 'pseudo' Nhanes data.









From the frequency, we can see that except for age, the other continuous variables looks close to normal. We will further testify their distribution later during the simulation part.

Methods

In order to follow the original transportability analysis, this report will follow the following general analytically design:

1. Acquiring both source data and target data. The goal is to see how a candidate model, which is build based on the source data, would perform on the target data. The evaluation used is Brier Score
2. If the model is mis-specified, modify the model on the combined training dataset of source and target to get a tailored model. Then evaluate this model on combined test data set of source and target.

   The model has the following formula

$$\log\left(\frac{P(CVD)}{1-P(CVD)}\right) = \log(HDLC)+\log(TCHOL)+\log(AGE)+\log(SYSBP_{UT}+1)+\log(SYSBP_T+1)+SMK+DIB$$

3. For generating tailored model, first estimate the probability of source population membership using training data form the source population. Second use the estimated probability to construct inverse-odds weights for the same observations. Lastly, apply the inverse-odds weights to estimate a tailor model using all source and training dataset observation.
4. The final step is to evaluate this tailored model on the pre-separated test data sets using the Brier score Estimator as follows:

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

   where

$$\hat{o} = \frac{\Pr\left[S = 0 \mid X, D_{\text{test}} = 1\right]}{\Pr\left[S = 1 \mid X, D_{\text{test}} = 1\right]}$$

Starting form train-test split, since there's missingness in the target Nhanes data, we will use mice function(Buuren and Groothuis-Oudshoorn 2011) to impute the missing data and also incorporate training test split process. By the end of MICE stage, we will have 5 unique train and test sets from target population with 75% vs 25% proportion. Next, for each of the train data of target, we will combine it with the same proportion splitted train set of source population to get a combined training set that is complete and ready for model tailoring. We will conduct the process for both men and women splitted subsets since the model should be evaluated on through such stratification and thus also tailored on such stratification.

In conclusion, with our 5 time imputation, we will end up with two lists of brier scores, one for men and the other for women. In each list, it consists of 5 different estimators of brier scores, which represent corresponding tailored model from combined training data, evaluated on combined test data. We will use average of each list as our final results for the transportability analysis.

Model Evaluation

The Brier Score measures the accuracy of predicted probabilities for binary outcomes. It ranges from 0 to 1, with lower values indicating better predictive accuracy. We can observe relative consistency of output across imputation for both male and female. The average brier score for male is about 0.1307 and that of women is 0.0557. While both value are very close to 0, indicating good predictive accuracy. This translates to the following conclusion:

Table 2: Variables with missingness

	Missing Percentage
SYSBP	15.7018443
HDLC	10.3227459
TOTCHOL	10.3227459
BPMEDS	6.1475410
BMI	5.7120902
DIABETES	0.0256148

Through transportability analysis, the prediction model derived based on Framingham data is evaluated to be also perform very well in the Nhanes data. Predictive accuracy of women subset is better than that of men subset.

Table 3: Brier Score results for Men

M1	M2	M3	M4	M5	Mean
0.1244	0.1241	0.1243	0.1241	0.1243	0.1242

Table 4: Brier Score results for Women

M1	M2	M3	M4	M5	Mean
0.0592	0.0596	0.0592	0.0592	0.0592	0.0593

Simulation

We will simulate individual level data from the summary of Nhanes data with insights gained from Framingham data, which is the source data.

The aim of this simulation study is to test the transportability of CVD-prediction model generated on the Framingham data to target population in different distributions(Similarity of target population to source population). We will use several different data generation mechanisms to generate individual level data of Nhanes. The different data generation mechanisms will incorporate situations where the simulated data is very close to distribution of the source population and not close to it. For each data generation mechanisms, we will run 5000 simulations using the same method from the above, and report average brier score for each scenario of data generation. The reason for choosing number of simulation to be 5,000 was based on the following reasons: referring to the objectives of the study, we want to how each estimators compare to the true value of brier score and a similar example provided in the reference paper used 10,000 number of simulations. In consideration of computation time and higher model complexity in comparison to the reference paper simulation example, we decided to use 5,000 number of simulations. We will also compare the estimators to the man and women brier score from the actual data as the true estimands. The performance measures is the respective averaged brier score from each of the different data generating process. It will be compared to non-simulated Nhanes dataset which corresponds to the results from the upper section. Seed for simulation experiments is set in the very beginning of this section as 2550 for reproducibility.

To start, we present this logged summary statistics from Nhanes data.

The reason for representing logged summary is that we make assumptions in the first data generation mechanism that all continuous variables follow a normal distribution after log transformation. This is a rather strong assumption and also the assumption *A1: Independence of the outcome Y and the population S, conditional on covariates from the paper* (Steingrimsson et al. 2022). Therefore in the first data generation,

Table 5: Logged Variable Summary of Nhanes Data

Variables	Summary
TOTCHOL	5.24(0.21)
AGE	3.93(0.24)
SYSBP	4.83(0.14)
HDLC	3.93(0.28)
BMI	3.39(0.23)
BPMEDS	1193(33%)
DIABETES	644(17%)
SEX	2015(52%)
CURSMOKE	787(20%)

we will generate each continuous variable based solely on the mean and sd from the above table and each binary variable as randomly generated number following binomial distribution with proportion of from the table as probability parameter. We run the simulation 5000 times, each simulation we generate 4000 cases of individual level data to match the actual Nhanes data, and get the average result for both men and women subset.

In the second data generation mechanism, we will use information from Framingham data to inform simulation of Nhanes data. By assuming multivariate normal distribution of all continuous variables, we will generate all the continuous variable of Nhanes data by using means from log transformed Nhanes original data and covariance matrix from log transformed Framingham data. For binary variable, since the proportion of each level for variable 'BPMEDS' and 'DIABETES' is very imbalanced in Framingham data, indicated form the above summary, we will generate them as usual of binomial distribution. For 'CURSMOKE' and 'SEX', we prefitted a logistic regression of each against the remaining variables using log transformed Framingham data. During the simulation stage, we will use these models to generate the variables for Nhanes data. This way, we hope to catch relationships and correlations between variables within the Framingham data, and use such information to help simulation of Nhanes data. We run the simulation 5000 times, each simulation we generate 4000 cases of individual level data to match the actual Nhanes data, and get the average result for both men and women subset.

For the third data generation, we will exploit the package 'fitdistrplus'(Delignette-Muller and Dutang 2015) to find the best distribution. This tool helped us compare different distributions and select the one that matched our data the closest, based on statistical tests (AIC) and visual plots. The distribution we chose and its parameters were crucial in creating a realistic synthetic dataset for our study on the Framingham risk score model's applicability to different populations.

We provide four candidate distribution: normal, exponential, gamma and log normal. Then we would fit each continuous variables withe the four candidates and select the best fit by lowest AIC values. The result is displayed as below. All variables except 'HDLC' is determined to be log normal distribution, and HDLC is best selected as gamma distribution. Following this, we will make the third data generation process follow their respective distribution and parameters. All binary variables are generated as the first generation. We run the simulation 5000 times, each simulation we generate 4000 cases of individual level data to match the actual Nhanes data, and get the average result for both men and women subset.

Table 6: Best Distribution fit for Continuous variables

	TOTCHOL	AGE	HDLC	BMI	SYSBP
Distribution	lnorm	lnorm	gamma	lnorm	lnorm
meanlog_shape	5.4533	4.0711	10.3745	3.2427	4.9185
sdlog_rate	0.1859	0.1245	0.2114	0.1466	0.1563

Finally, we compile all the results together into this table. The true result is represented as the result we get

first hand from the individual level data of Nhanes. Through comparison, we can see that data generation 1 has the closest model performance in terms of transportability analysis with the true result. The third generation mechanism, which includes testing of univariate distribution, has relative large value of brier scores. This means the model trained on Framingham data generate poorly on the data simulated under this mechanism. Then we make comments on the simulation bias and standard errors:

1. **Type 1 Data Generation**: This type shows a bias of -0.0191 for man and 0.0073 for women , indicating a slight systematic deviation from the true value. The standard error suggests low variability in the results across different simulation runs.
2. **Type 2 Data Generation**: For this category, the bias of 0.0346 for man and -0.025 for women, which means the simulation results are quite consistent over the true value to. The standard error reflects greater consistency in the simulation outputs compared to Type 1.
3. **Type 3 Data Generation**: This type has a bias of 0.1488 for man and 0.1468 for women, showing the most significant deviation from the true value among the three types. Its standard error, indicates the biggest spread in results.

In each case, the bias and SE collectively provide insight into the accuracy and reliability of the simulations under different data generation scenarios. Comparing these metrics across the three types helps identify which method aligns closest with the true values and offers the most consistent results.

Table 7: Average Brier Score Comparison between True and differnt data generation

	True	Gen_1	Gen_2	Gen_3
Men	0.1242	0.1051(0.000237)	0.1588(0.000409)	0.2729(0.000273)
Women	0.0593	0.0666(0.000206)	0.0342(0.000114)	0.2062(0.000426)

Note:
Result is presented as Mean(SE)

Table 8: Simulation Bias Comparison between differnt data generation

	Gen_1	Gen_2	Gen_3
Men	-0.0191(4e-06)	0.0346(7e-06)	0.1487(3e-05)
Women	0.0073(2e-06)	-0.025(5e-06)	0.1469(3e-05)

Note:
Result is presented as Bias(SE)

Conclusion and Discussion

The transportability of the CVD prediction model from the Framingham source population to the NHANES target population, as evidenced by our simulation results showing low bias and low standard error, marks a significant advancement in predictive modeling. The model's commendable predictive accuracy in both male and female subsets of the NHANES data not only underscores its potential for broader application but also aligns with our study's goal of assessing model applicability across diverse populations. This gender-based differentiation in predictive accuracy further stresses the importance of considering sex-specific variations in CVD risk factors, enriching our understanding of how prediction models can be tailored for individual subgroups within different populations.

However, this analysis does bear certain limitations. Primarily, the assumptions of representativeness between the source and target populations, and the methodological simplifications in our simulation study, may

not fully address all demographic and lifestyle differences between the Framingham and NHANES cohorts. In line with the reference paper, the assumptions of the independence of outcome Y and the population S conditional on covariates, and positivity, are critical. Our findings, particularly in the simulation segment demonstrating the significant effect of different target data distribution assumptions on the performance of the source-derived model, echo this sentiment. This indicates the necessity for model customization when applying it to the target data.

Additionally, while the Brier Score served as a primary metric for evaluation, a more comprehensive analysis incorporating additional performance metrics such as AUC and ROC curves, especially in classification tasks, could provide a broader perspective on the model's transportability. Nonetheless, it is essential to adapt these estimators to the unique context where outcome data in the source variable may be limited or absent.

In conclusion, these findings, promising as they are, highlight the need for continued research to further refine and validate the model's performance across diverse, real-world target populations. This includes employing a wider array of evaluation measures and considering potential changes over time in the target populations, thereby aligning with the initial objectives set forth in our report.

Reference

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "{Mice}: Multivariate Imputation by Chained Equations in r" 45: 1–67. https://doi.org/10.18637/jss.v045.i03.

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." *Circulation* 117 (6): 743–53. https://doi.org/10.1161/circulationaha.107.699579.

Delignette-Muller, Marie Laure, and Christophe Dutang. 2015. "{Fitdistrplus}: An {r} Package for Fitting Distributions" 64. https://doi.org/10.18637/jss.v064.i04.

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.