Tranportability Case-study and Simulation Analysis

Yu Yan

2023-11-29

Abstract

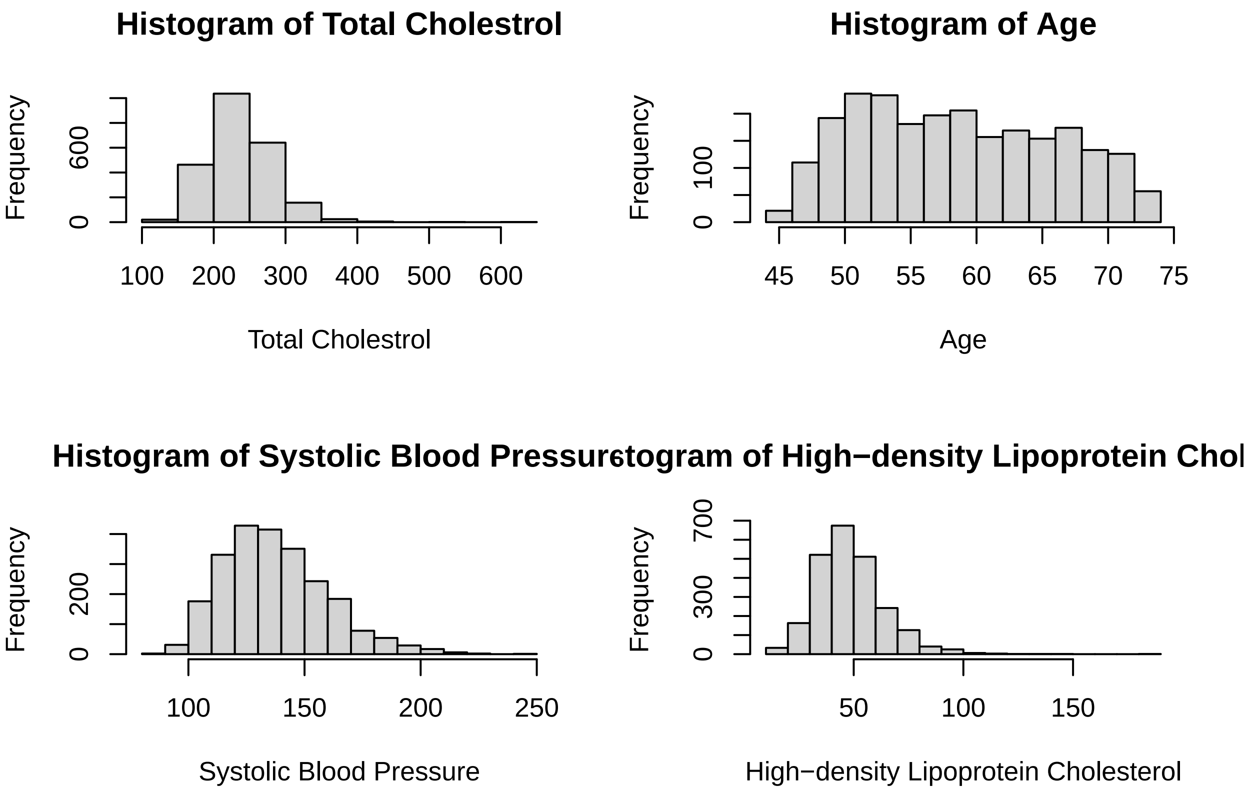
This report focuses on a case_study of Transportability analysis following the methodology outlined in the reference paper titled "Transporting a Prediction Model for Use in a New Target Population" by Dr. Steingrimsson et.al. It involved using source data from Framingham Heart Study, target data from nhanes study and a model derived from source data. The first goal of this analysis report is to evaluate the performance of the given prediction model in the target population underlying the NHANES data.

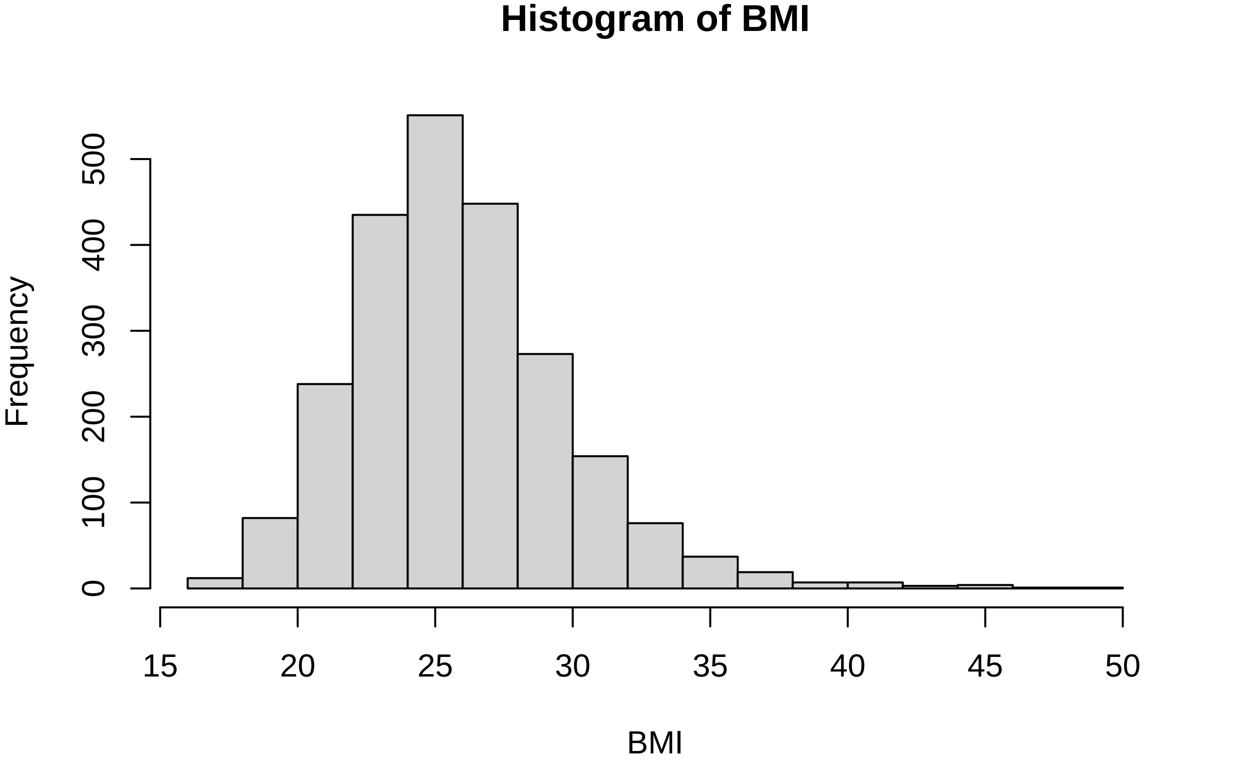
The transportability analysis evaluates model performance using the Brier Score and involves model modification on combined training data, with tailored models constructed using inverse-odds weights. By evaluting on combined test data, the results indicate high predictive accuracy in both male and female subsets, with women outperforming men.

The report also briefly touches on a simulation study, conducted to test the transportability of the CVD prediction model across different data generation mechanisms. The goal is to simulate individual level data as if only summary level data is presented. The case illustrated using nhanes data. Three mechanisms are explored, each simulating NHANES data under varying assumptions. Results reveal the impact of these mechanisms on model performance, underlining the importance of considering diverse data generation scenarios when assessing model transportability in new target populations.

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.





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:

   $$\log\left(\frac{P(CVD)}{1-P(CVD)}\right) = \log(HDLC){+}\log(TOTCHOL){+}\log(AGE){+}\log(SYSBP_{UT}{+}1){+}\log(SYSBP_{T}{+}1){+}CURSMC$$

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-seperated test data sets using the Brier score Estimator as follows:
   $\hat{\psi}_{\hat{\beta}} = \frac{\sum_{i=1}^{n} I(S_i{=}1, D_{\text{test }, i}{=}1)\hat{o}(X_i)\left(Y_i - g_{\hat{\beta}}(X_i)\right)^2}{\sum_{i=1}^{n} I(S_i{=}0, D_{\text{test }, i}{=}1)}$
   where $\hat{o} = \frac{\Pr[S{=}0|X, D_{\text{test }}{=}1]}{\Pr[S{=}1|X, D_{\text{test }}{=}1]}$

Starting form train-test split, since there's missingness in the target nhanes data, we will use mice function 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.

Table 1: Variables with missingness

	Missing Percentage
BMI	5.7120902
HDLC	10.3227459
BPMEDS	6.1475410
TOTCHOL	10.3227459
DIABETES	0.0256148
SYSBP_UT	15.8043033
SYSBP_T	11.0655738

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:

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

Table 2: Brier Score results for Men

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

Table 3: 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 nhances data with insights gained from Framingham data, which is the source data.

The aim of this simulation study is to test the Tranportability 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 nhance. 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.

The following are two tables of logged summary statistics of both complete case nhances and Framingham 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 covariatesm from the paper* (Steingrimsson et al. 2022). Therefore in the first data generation, 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

Table 4: Summary Statistics of log tranformed Framingham data

Characteristic	**N = 2,348**
TOTCHOL	5.46 (5.33, 5.58)
AGE	4.08 (3.97, 4.17)
SYSBP	4.91 (4.80, 5.02)
HDLC	3.87 (3.66, 4.04)
BMI	3.24 (3.15, 3.33)
BPMEDS	
0	2,008 (86%)
1	340 (14%)
DIABETES	
0	2,178 (93%)
1	170 (7.2%)
SEX	
0	1,022 (44%)
1	1,326 (56%)
CURSMOKE	
0	1,498 (64%)
1	850 (36%)

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

Table 5: Summary Statistics of log tranformed Nhanes data

Characteristic	**N = 3,904**
TOTCHOL	5.24 (5.11, 5.38)
Unknown	403
AGE	3.99 (3.74, 4.14)
SYSBP	4.84 (4.74, 4.91)
Unknown	613
HDLC	3.91 (3.71, 4.11)
Unknown	403
BMI	3.37 (3.23, 3.53)
Unknown	223
BPMEDS	
0	2,471 (67%)
1	1,193 (33%)
Unknown	240
DIABETES	
0	3,259 (83%)
1	644 (17%)
Unknown	1
SEX	
0	1,889 (48%)
1	2,015 (52%)
CURSMOKE	
0	3,117 (80%)
1	787 (20%)

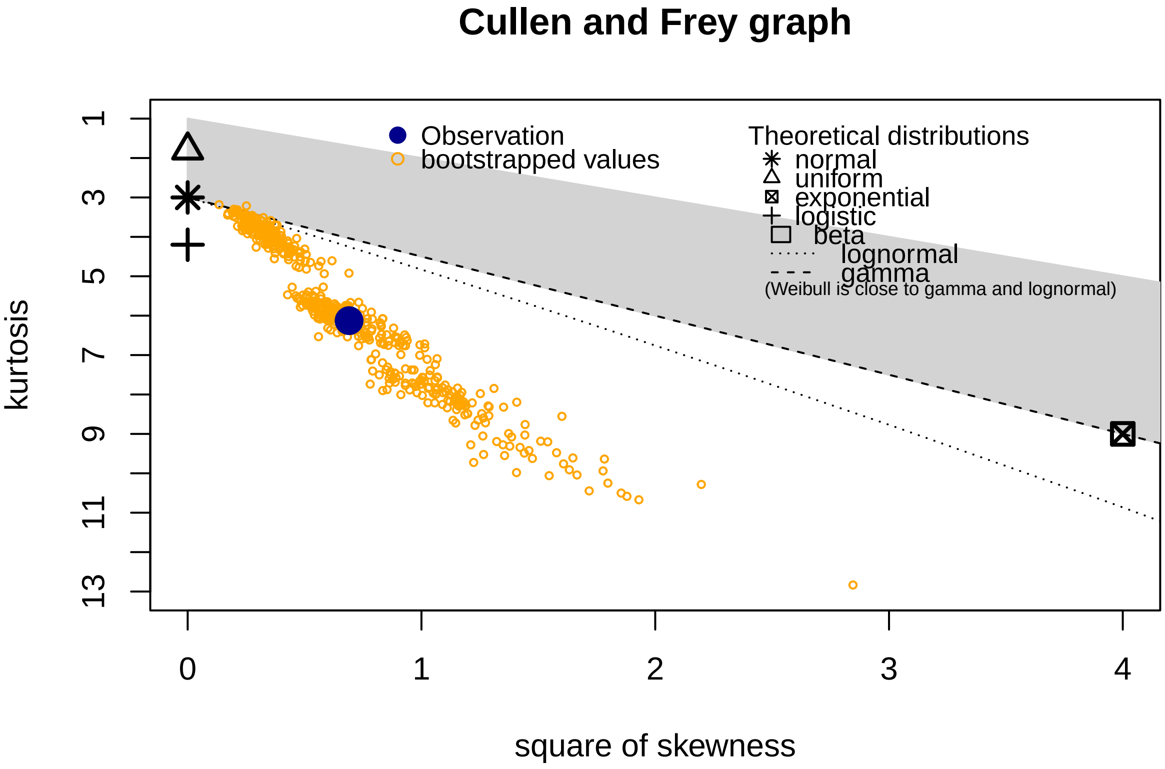
[1] Median (IQR); n (%)

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 nanes data. This way, we hope to catch relationships and correlarions 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 that fits for each continuous variable and get their parameter values, if we don't make assumption about all being normal distribution.

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. We can also draw the Cullen and Frey graph to see which distribution best fits the data. Example of graph is given below. 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.



```
## summary statistics
## ------
## min:  130   max:  625
## median:  235
## mean:  237.6052
## estimated sd:  44.96825
## estimated skewness:  0.8308584
## estimated kurtosis:  6.126716
```

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 repersented 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.

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

	True	Gen_1	Gen_2	Gen_3
Men	0.1242083	0.1740712	0.2870589	0.4725363
Women	0.0592728	0.0665538	0.0342457	0.2060685

Conclusion and Discussion

The successful transportability of the CVD prediction model from the Framingham source population to the NHANES target population is a noteworthy application. The model's high predictive accuracy in both male and female subsets of the NHANES data indicates its potential applicability across diverse populations, offering promise for its real-world utility. The gender-specific analysis also highlights the importance of considering sex-based differences in CVD risk factors, providing valuable insights for tailoring prediction models to specific subgroups within target populations.

However, it's important to acknowledge certain limitations in this analysis. The assumption of source population representatives and the simplifications made in the simulation study might not capture all demographic and lifestyle variations between the Framingham and NHANES datasets. As is also mentioned in the refernce paper the two assumptions: Independence of the outcome Y and the population S, conditional on covariates; and positivity. This can also be partly illustrated in the simulation part that varying distribution assumption of the target data could impact greatly the performance of model that is derived on source population, even though the model gets tailored before applied to the target data.

Furthermore, the sole reliance on the Brier Score as an evaluation metric, without considering additional performance metrics or external validation measures, could provide a limited perspective on model transportability.The report could incorporate other predicitvce measures such as AUC and ROC curves for classification tasks. Only to note the estimator must be tailored and account for the fact that there's potentially no outcome data in the source variable.

In conclusion, while these findings are promising, they underscore the need for further research to refine and validate the model's performance in real-world scenarios with diverse target populations, considering a broader range of evaluation measures and accounting for potential temporal changes.

Reference

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.