Project 3

Han Ji

2023-12-05

Abstract

Evaluating the transportability and the performance is important when applying a prediction model derived in the source population to the target population, especially when the characteristics of two populations differ. Recently, a method that uses inverse-odds weights to get transported brier score estimate has been developed. In this study, we first examine its performance when transporting a prediction model for CVD risk on Framingham Study to NHANES data, while the population in NHANES seem to be healthier than population in Framingham. The average brier score estimated from imputation are 0.0929 and 0.0563 in NHANES for men and women, compared to 0.1939 and 0.1171 for Framingham. We also conduct a simulation study to evaluate the transportability when only summary statistics of the target population are available. We use two methods, multivariate normal distribution and bootstrapping from the source population, to simulate the individual-level data based on the summary statistics of NHANES. The simulated data from the first method is well-calibrated for the mean statistics, and the other is not calibrated. Multivariate normal has a smaller relative bias (13.47% versus 38.93%) for men in brier scores, while bootstrap is slight better for women (-1.62% versus -1.46%). We conclude that the difference in the oracle and the simulated brier score seems to be associated with the difference in the source and the target brier score, and this impacts differ for different simulation methods.

Introduction

Besides the prediction accuracy, the ability to generalize is also an important aspect for prediction models. For example, a health-care system may want to apply a prediction model developed from another health-care system to their own patients. However, a well-performed model may not have a good performance when applied to another population.

Recently, a method that can estimates the performance of transporting a prediction model in a new target population has been published (Steingrimsson et al. 2023). This methods uses inverse-odds weights based on the covariates to account for the differences in the characteristic distributions of the source and the tartget population. It doesn't require the outcome to be available in the target population, and it provides an estimate of brier scores by a weighted average of the brier scores components from the source population.

In this study, we implement this method to transport the prediction model from Framingham Heart Study to the National Health and Nutrition Examination Survey (NHANES) data. In addition, we consider a realistic scenario that only summary statistics from NHANES data is available. We simulate individual data based on the summary statistics, and we evaluate the performance of transporting the prediction model compared to the oracle estiamte when the individual-level data is available.

Methods

Framingham Heart study

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. This dataset in our study is the teaching dataset from the Framingham Heart Study (No. N01-HC-25195), provided with permission from the National Heart, Lung, and Blood Institute (NHLBI). The Framingham Heart Study is conducted and supported by the NHLBI in collaboration with Boston University (Grembi 2022). The original eligibility for age is from 30 to 74 (D'Agostino et al. 2008). After removing missing values, 2539 observations are considered in this study (n = 1094 for men; n = 1445 for women). We included the outcome variable CVD, covariate sex, total cholesterol, age, systolic blood pressure, smoking status, diabetes, blood pressure medication, high density lipoprotein cholesterol, and BMI. We create two new variables, SYSBP_UT and SYSBP_T, to differentiate the systolic blood pressure under the treatment and without the treatment. If under treatment, SYSBP_T will be equal to the measured systolic blood pressure, while SYSBP_T will be equal to 0, and vice versa.

National Health and Nutrition Examination Survey

National Health and Nutrition Examination Survey (NHANES) is a survey conducted by the National Center for Health Statistics (NCHS), and data are publicly available at: https://www.cdc.gov/nchs/nhanes.htm. We select the data from 2017 and 2018 and the variables present in the Framingham study model (Endres 2023). There are 9254 observations at the beginning, and we remove the observations that don't match the age eligibility (<30 yrs) and the ones with more than 3 variables missing. There are 3726 observations left (n = 1795 for men; n = 1931 for women).

Multiple imputation

We first train-test split both Framingham and NHANES data by 3:1 ratio for each sex, so the ratio between two studies are the same in train and test set. We use `mice` to conduct multiple imputation with in the NHANES train set data for each sex, and we apply the same model to the NHANES test set data. The seed used is 2550. The full train and test set is generated by combing Framinghan and NHANES data for each sex. The population indicator S is attached: S = 1 if in Framingham study and S = 0 if in NHANES.

Prediction Model

The prediction model for CVD risk is based on previous studies (D'Agostino et al. 2008). We fit the logistic regression model on the Framingham train set for each sex and apply to the test set.

Transportability Analysis

To avaluate the prediction performance on the target population NHANES, we use the brier score formula below (Steingrimsson et al. 2023):

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

where $\hat{o}(X_i)$ is an estimator for the inverse-odds weights in the test set, $\frac{Pr(S=0|X,D=1)}{Pr(S=1|X,D=1)}$.

Simulation

We simulate the individual-level data based on the summary statistics of the target population. We use ADEMP framework to design our simulation studies:

Aims The aim of this simulation study is to evaluate the performance of simualted data sets from summary statistics of the target population during transporability analysis. Our main consideration includes: 1) the impact of the difference between the source and target population on the performance; 2) the assumption on the target population distribution derived from the source population individual data; 3) the calibration of the simualted data set compared to the given summary statistics

Data-generating Mechanisms: We consider two mechanisms:

1. Multivariate normal. Since the marginal distribution of all continuous variables in Framingham study looks normal after log, we use a multivariate normal distribution to generate samples. The mean is the log mean values from NHANES, while the covariance matrix is from Framingham since the standard deviation is hard to estimate after log. The continuous variable values are obtained by taking exponential of the simulated values, and the binary variable values are obtained by using quantiles of the simulated values and the observed proportions from NHANES.

2. Bootstrap. We use normal density for continuous variables and the proportion for categorical variables using the mean and sd from NHANES to get a joint density estimate for each observation in Framingham. It represents the likelihood that this observation can belong to the NHANES study. After scaling the density to probability by dividing by the sum, we boostrap the observations from Framingham to generate a sample for the tartget population.

Estimand The estimand is the Brier score in the target population used in the transporability analysis.

Methods The same as the previous analysis on the actual NHANES data, we train-test split the Framingham and the simulated NHANES data, and we calculate the transported brier scores. Due to the complexity, we repeat the simulation 1000 times without estimating the required simulation number by Monte Carlo SE.

Performance measures We consider the bias, relative bias, and MSE for both methods and each sex compared to the oracle estimate when individual values are available.

Results

Table 1: Summary of characteristics stratified by sex in the Framingham study.

	Men	Women	p
n	1094	1445	
CVD = 1 (%)	360 (32.9)	242 (16.7)	<0.001
TOTCHOL (mean (SD))	226.44 (41.49)	246.32 (45.51)	<0.001
AGE (mean (SD))	60.01 (8.18)	60.55 (8.40)	0.106
SYSBP (mean (SD))	138.94 (20.89)	139.94 (23.71)	0.272
CURSMOKE = 1 (%)	425 (38.8)	445 (30.8)	<0.001
DIABETES = 1 (%)	96 (8.8)	95 (6.6)	0.045
BPMEDS = 1 (%)	123 (11.2)	259 (17.9)	<0.001
HDLC (mean (SD))	43.63 (13.37)	53.07 (15.67)	<0.001
BMI (mean (SD))	26.25 (3.47)	25.55 (4.22)	<0.001

Table 2: Summary of characteristics stratified by sex in the NHANES study.

	Men	Women	p
n	1795	1931	
TOTCHOL (mean (SD))	188.54 (41.44)	196.10 (40.01)	<0.001
AGE (mean (SD))	53.07 (12.20)	52.00 (11.99)	0.007
SYSBP (mean (SD))	128.27 (17.70)	125.95 (20.29)	<0.001
CURSMOKE = 1 (%)	441 (24.6)	317 (16.4)	<0.001
DIABETES = 1 (%)	339 (18.9)	283 (14.7)	0.001
BPMEDS = 1 (%)	561 (33.5)	585 (32.0)	0.392
HDLC (mean (SD))	47.76 (14.07)	57.52 (15.83)	<0.001
BMI (mean (SD))	29.90 (6.55)	30.91 (8.19)	<0.001

Table 3: Model coefficients of logistic regression for cardiovascular disease on the full Framingham study data stratified by sex.

	Men	Women
log(HDLC)	-0.661	-1.224
log(TOTCHOL)	1.031	0.895
log(AGE)	4.793	5.089
log(SYSBP_UT + 1)	1.874	2.419
log(SYSBP_T + 1)	1.947	2.486
CURSMOKE	0.264	0.518
DIABETES	0.758	1.086

Table 1 summarizes the characteristics of the Framingham study stratified by sex, and it shows that the distributions differ by sex for the most of variables. Women participated in the study had a better health level as they had a lower total cholesterol, a lower percentage of smoking, and a lower BMI. Correspondingly, the proportion of being diagnosed with CVD is higher in men than women. Since these two subpopulations seem to be different, we separate them for the following analysis, though some covaraites may associate with the outcome in a similar way between women and men.

We fit a logistic regression model on the subset for each sex, and the model coefficients are shown in Table 2. All continuous variables are transformed to the natural log scale because. All variables included are significant ($p < 0.05$) except smoking status in men ($p = 0.07$). The overall trend in the coefficients is similar between men and women but there are still subtle differences: the smoking status is more important for women than men. For women, if the increase in total cholesterol is mainly HDLC, it is still beneficial for CVD risk; however, for men, any increase in total cholesterol is harmful on the population level.

Now we want to transport the prediction model from the Framingham study to the NHANES study, while the outcome variable is missing (Table 2). We first train-test split the Framninghan and NHANES data seperately for each sex with the ratio 3:1, so the ratio between two studies are the same between the train and test set for each sex. Multiple imputation is conducted to generate 5 complete set for train and test set, and the final brier scores are calculated for each imputation.

Table 4: Brier scores estimated on the target population, NHANES, stratified by sex from each imputation. The last row is the brier score calculated from the source population, Framingham study.

Imputation	Men	Women
1	0.0931	0.0566
2	0.0930	0.0564
3	0.0937	0.0570
4	0.0928	0.0567
5	0.0921	0.0550
Avg.	0.0929	0.0563
Source	0.1939	0.1171

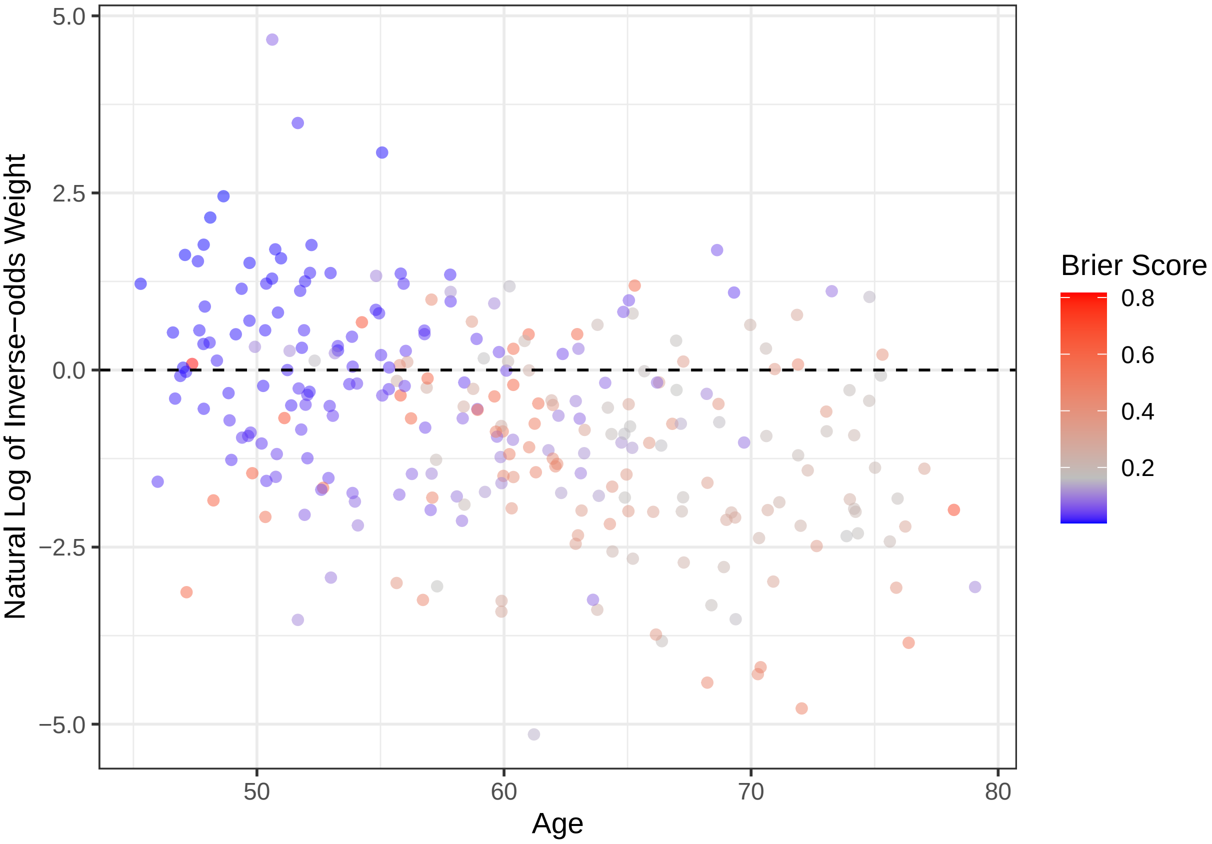

Figure 1: Log of inverse-odds weights of the framingham data in the test set vesus age, colored by the unweighted brier score based on the prediction model from the train set. The grey color represents the mean brier score for all observations, and the black dashed line represents inverse-odds weight equal to 1.

Table 4 shows the summary table of brier scores. The average for NHANES is 0.0929 and 0.0563 for men and women, respectively, and both are lower than Framingham, 0.1939 and 0.1171. The transported brier scores vary among multiple imputations but all are close the mean values, meaning the inverse-odds weight estimates tend to be stable overall. We also look at what causes the obvious decrease in the brier scores. We plot the inverse-odds weight, age, and the brier score components of each observation from the framingham test set based on one imputation result (Figure 1). The points tend to have a higher brier score when age increases, meaning the prediction accuracy decreases when age increases. On the other hand, we see the overall inverse-odds weight decreases when age increases, meaning the effect of the corresponding component is less on the brier score as age increases. The ones with a brier score lower than average has a weight higher than 1, and the ones with a higher brier score has a smaller weight, and thus the resulting brier score is lower in NHANES. This means that the observations with a smaller age have a higher prediction accuracy, which implies the health level may be associated with the prediction accuracy. This is also reflected in the summary characteristics that the population in NHANES is more healthy than Framingham overall. However, this trend is not obvious in other predictors. Another reason could be that the prediction on the younger population is easier given other covaraites, while it is more complex for the elder population.

We then test for the case when only the summary statistics for the target population is available but the individual data cannot be accessed. Following the ADEMP framework, we design two simulation methods to generate the individual level for the target population data (Morris et al. 2019). The first one is using a multivariate normal distribution to generate data. Based on our EDA (not shown due to length limit), the marginal distribution of each continuous covariates follows a normal distribution approximately after taking natural log. However, the standard deviation of the distribution becomes different after log, and it is hard to estimate. Since the standard deviations of Framingham and NHANES study are at the same scale, we use the log of the mean from NHANES study and the covariance matrix from Framingham study after log to generate the multivariate normal distribution. The second method is using bootstrapping to generate samples based on the observations from Framingham study. If the covariates are close to the mean statistics of NHANES, this observation has a higher probability being drawn.

Table 5: Summary statistics of the simulated target population for men (empirical SD in parentheses).

	Multivariate Normal		Bootstrap		Actual	
	Mean	SD	Mean	SD	Mean	SD
TOTCHOL	191.68 (0.82)	35.18 (0.66)	207.69 (0.6)	26.7 (0.45)	188.54	41.44
AGE	53.56 (0.17)	7.29 (0.13)	56.87 (0.16)	6.71 (0.11)	53.07	12.2
SYSBP	129.65 (0.44)	19.08 (0.34)	130.03 (0.28)	12.1 (0.21)	128.27	17.7
CURSMOKE	0.25 (0)	NA	0.19 (0.01)	NA	0.25	NA
DIABETES	0.19 (0)	NA	0.01 (0)	NA	0.19	NA
BPMEDS	0.33 (0)	NA	0.03 (0)	NA	0.33	NA
HDLC	50.13 (0.37)	15.94 (0.37)	43.95 (0.21)	8.84 (0.17)	47.76	14.07
BMI	30.16 (0.1)	4 (0.07)	26.98 (0.07)	2.94 (0.06)	29.9	6.55

Table 6: Summary statistics of the simulated target population for women (empirical SD in parentheses).

	Multivariate Normal		Bootstrap		Actual	
	Mean	SD	Mean	SD	Mean	SD
TOTCHOL	199.39 (0.83)	36.29 (0.69)	217.18 (0.65)	27.02 (0.43)	196.1	40.01
AGE	52.5 (0.16)	7.29 (0.13)	55.83 (0.15)	6.43 (0.12)	52	11.99
SYSBP	127.71 (0.49)	21.31 (0.39)	126.98 (0.29)	13.19 (0.24)	125.95	20.29
CURSMOKE	0.16 (0)	NA	0.1 (0.01)	NA	0.16	NA
DIABETES	0.15 (0)	NA	0.01 (0)	NA	0.15	NA
BPMEDS	0.32 (0)	NA	0.04 (0)	NA	0.32	NA
HDLC	60.04 (0.42)	17.94 (0.39)	53.6 (0.23)	10.42 (0.17)	57.52	15.83
BMI	31.3 (0.11)	4.93 (0.09)	25.87 (0.08)	3.69 (0.07)	30.91	8.19

For calibration, we compares the distribution of the summary statistics for both methods after 1000 simulations with NHANES (Table 5 and 6). We see the mean values from multivariate normal method are pretty close to the values given, especially for binary variables since the values are specified based on quantiles. However, the standard deviation is deviated probably because we use the covariance matrix from Framingham for approximation. The bootstrap method has a worse calibration for both mean and standard deviation. Standard deviation is underestiamted tha the actual NHANES data probably because we bootstrap samples from Framingham.

Table 7: Peformance of brier scores in the target population for multivaraite normal and bootstrap methods (Monte Carlo SEs in parentheses).

	Men		Women	
	Multivariate Normal	Bootstrap	Multivariate Normal	Bootstrap
Bias	0.01252 (0.00031)	0.03618 (0.00015)	-0.00091 (0.00021)	-0.00082 (0.0001)
Relative Bias	13.47%	38.93%	-1.62%	-1.46%
MSE	0.00025 (0.00001)	0.00133 (0.00001)	0.00004 (<0.00001)	0.00001 (<0.00001)

Table 7 shows the performance metrics of two methods for men and women, and relative bias is used since brier score is pretty small. The mean brier scores of multivariate normal method are 0.10547 and 0.05544 for men and women, and the ones of bootstrap is 0.12912 and 0.05553 for men and women. Compared to the oracle estimate from NHANES data, the bias is more than 10% for both methods in men, and the bootstrap has a higher bias and MSE than multivariate normal while it has a smaller Monte Carlo SE. For women, both methods perform well while bootstrap has a slightly smaller bias and Monte Carlo SE, and thus a smaller MSE.

Discussion

In this study, we first estimate the performance of the prediction model on Framingham to NHANES where the outcome variable is unknown. Although the populations from both studies come from the United States, the characteristics differ a lot, meaning we cannot directly transport our prediction model from Framingham to NHANES. By using inverse-odds weights based on the covariates in the test set, we can get an estimate of brier score and evaluate the performance of this prediction model on NHANES if the outcome CVD is available.

We then consider a more realistic setting that many studies only provide the summary statistics of the study population without individual-level data due to privacy and other reasons. We conduct a simulation study to compare two methods that generate samples based on summary statistics. The multivariate normal method assumes that the distribution of the source and target population should both look normal after log transformation; the bootstrap method is similar to the idea of weighting that observations from the source population that resemble the target population are drawn, and it doesn't require a distribution assumption. Here we use normal density with the target population mean and standard deviation for simplicity, but it could be other forms as long as that the observations closer to the target summary statistics are likely to be drawn.

The first method use the covariance matrix from Framingham to preserve the association among variables, but it could violate positivity assumption as the generated values could be out of the scope of the source or the target population. This method has a great calibration on mean values but not standard deviation. This could be because the underlying distributions of covariates are different. For example, age follows a uniform distribution approximately in the actual NHANES data. The second method directly draws values from Framingham to not only maintain the association among variables but also doesn't violate positivity assumption, but it has a poor calibration on mean since the two population differ a lot in summary statistics. It also has a lower variability than the actual NHANES data due to over-representing observations in bootstrapping.

For performance, the multivariate normal method is overall better than bootstrap, especially for men. This is probably due to the huge difference in brier score estimated between Framingham and NHANES, while the multivariate normal has a good calibration. Bootstrap is slightly better in women, and the difference in brier score is smaller in women between Framingham and NHANES. Potentially, boostrap could be better than multivariate normal if the difference between the oracle estimate of the target population and the source population. However, this cannot be testable, and it requires some reasoning based on the summary statistics.

Our study has several limitations. First, we use multiple imputation to get the oracle estimate for the individual-level data in NHANES. This method could be inappropriate if the missing mechanism is missing not at random (MNAR). Second, we use the covariance from Framingham data, which doesn't use the standard deviation of the NHANES data. We may need to work on this method and figure out how to incorporate this information while still assuming the distribution is approximately normal after natural log. Third, our prediction and imputation model are separated by sex, but using both sex together may yield a more accurate association among variables.

For future direction, we can conduct a comprehensive simulation study that we simulates data with different underlying distributions instead of using real-world data. This helps us to test the performance of two methods under different scenarios. Second, we can consider using other forms of density instead of normal density for the bootstrap methods. It could change or even improve our results.

Conclusion

In our study, we estimate the performance if we transported a prediction model on Framingham data for CVD risk to the target population NHANES, where the outcome is not available. The data is train-test split and stratified by sex, and multiple imputation is conducted to generate complete data. The brier score is calculated for 5 imputed sets, and the average is 0.0929 and 0.0563 in NHANES for men and women, compared to 0.1939 and 0.1171 for Framingham. We also simulate NHANES data based on the summary statistics by multivariate normal and bootstrap methods. The multivariate normal performs better for men, while the bootstrap performs slightly better for women. The difference in the oracle and the simulated brier score seems to be associated with the difference in the source and the target brier score.

Code Availability

The code for analysis and generating this report is available on github: https://github.com/RuBBiT-hj/PHP2550_Project3

References

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. (2008), "General cardiovascular risk profile for use in primary care," *Circulation*, 117, 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579.

Endres, C. (2023), *nhanesA: NHANES data retrieval*.

Grembi, J. (2022), *riskCommunicator: G-computation to estimate interpretable epidemiological effects*.

Morris, T. P., White, I. R., and Crowther, M. J. (2019), "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, 38, 2074–2102. https://doi.org/https://doi.org/10.1002/sim.8086.

Steingrimsson, J. A., Gatsonis, C., Li, B., and Dahabreh, I. J. (2023), "Transporting a prediction model for use in a new target population," *Am. J. Epidemiol.*, Oxford University Press (OUP), 192, 296–304.