PHP2550_Project_3

Jialin Liu

2023-12-02

Introduction

Users of prediction models want to apply the models in a specific target population. Prediction models are often developed from samples in source populations, however, models cannot be directly applied to the target population since datasets are typically not random sample from the target population, even distributions of observed variables are totally different between source and target populations. Consequently, models built using the data from source population are not applicable to the target population so that model performance evaluation in the source population cannot perfectly reflect performance in the target population unless using tailored prediction models as an attractive alternative to evaluate performance in the target population to achieve transportability tasks. In many cases, both covariates and outcome are available in source populations, whereas only covariates are available in target populations without prior information about outcomes. Under the lack of outcomes in target populations, we tailor prediction models given outcomes information from the source population and assess performance of models for datasets with covariates only[1].

This project aimed to see if the prediction model can be generalized to the other target population by looking at the Brier scores from the transportability analysis, conducted a simulation study when individual level datasets are not available for transportability, and evaluated performance of Brier risk scores.

The Framingham Heart Study

It is widely accepted that age, sex, high blood pressure, smoking, dyslipidemia, and diabetes are the major risk factors for developing cardiovascular disease (CVD). The Framingham Heart Study was a landmark long term prospective study of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts, and identified effects of risk factors. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. The Framingham data has been used to create models for predicting cardiovascular risk given risk factors and markers of disease, such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, diagnoses of diabetes, and medication use[2]. From published scholar's work, the sex-specific multivariable risk factor algorithm was created to assess and predict CVD risk[3]. The study sample consisted of attendees of the baseline examinations free of prevalent CVD who were 30 to 74 years of age with non-missing data on covariates. Under this work's reference, we conducted variable selection including and applied models. After exclusions, 2438 participants (mean age, 59 years; 1380 women) remained eligible.

In Table 1, the risk factor characteristics of men and women in our sample at the baseline examinations are significantly different at the type-I error of 0.05. In our middle-aged sample, mean levels of systolic blood pressure and the prevalences of diabetes were similar in men and women. The prevalences of cigarette smoking and use of Anti-hypertensive medication were substantially higher in women. Then we created two new variables `SYSBP_UT` and `SYSBP_T` to get systolic blood pressure based on whether participants took medication or not. As we're not interested in measurement of hazard rates, we would like to remove censored data by examining risk within 15 years. Aiming to mimic models presented in the published works, we splitted the sample data by sex and fitted the sex-specific model with respect to log transforms for all continuous variables and selected categorical variables CURSMOKE and DIABETES to predict the probability of cardiovascular disease taking place.

Table 1: Characteristics of Risk Factors Stratified by SEX in the Framingham Data

	Men(1)	Women(2)	P-values
n	1058	1380	
CVD (mean (SD))	0.31 (0.46)	0.15 (0.36)	<0.001
TIMECVD (mean (SD))	7300.92 (2368.86)	8016.15 (1780.62)	<0.001
SEX = 2 (%)	0 (0.0)	1380 (100.0)	<0.001
TOTCHOL (mean (SD))	226.66 (41.56)	246.16 (45.96)	<0.001
AGE (mean (SD))	59.29 (7.60)	59.63 (7.65)	0.283
SYSBP (mean (SD))	138.48 (20.85)	139.03 (23.64)	0.548
CURSMOKE = 1 (%)	422 (39.9)	445 (32.2)	<0.001
DIABETES = 1 (%)	92 (8.7)	90 (6.5)	0.052
BPMEDS = 1 (%)	114 (10.8)	241 (17.5)	<0.001
HDLC (mean (SD))	43.61 (13.49)	53.17 (15.67)	<0.001
BMI (mean (SD))	26.28 (3.47)	25.55 (4.25)	<0.001

The National Health and Nutrition Examination Survey (NHANES)

Considering the Framingham data as the source population, we use the NHANES data from 2017-2018 with the same covariates as the target population. We select variables including systolic blood pressure BPXSY1, gender RIAGENDR, age RIDAGEYR, Body Mass Index BMXBMI, cigarette smoking SMQ040/SMQ020, serum total cholesterol LBXTC, HDL cholesterol LBXHDD, diabetes DIQ010, and use of Anti-hypertensive medication BPQ020/BPQ040A/BPQ050A. The we followed the same step proceeded in the source population to add two variables about systolic blood pressure with medication treatment or not, and to filter ages ranging from 30 to 74. After exclusions, 3189 participants (mean age, 52 years; 1632 women) remained eligible. In Table 2, mean levels of serum total cholesterol and HDL cholesterol in the target population at the baseline examinations were significantly higher in women. In our middle-aged sample, the prevalences of diabetes and cigarettes smoking, as well as mean levels of systolic blood pressure, were substantially higher in men.

Table 2: Characteristics of Risk Factors Stratified by SEX in the NHANES Data

	Men(1)	Women(2)	P-values
n	1557	1632	
SYSBP (mean (SD))	128.08 (17.12)	125.46 (19.88)	<0.001
SEX = 2 (%)	0 (0.0)	1632 (100.0)	<0.001
AGE (mean (SD))	52.96 (12.65)	51.88 (12.54)	0.015
BMI (mean (SD))	29.89 (6.27)	30.72 (8.20)	0.001
HDLC (mean (SD))	47.65 (13.95)	57.89 (15.96)	<0.001
CURSMOKE = 1 (%)	372 (23.9)	269 (16.5)	<0.001
BPMEDS = 1 (%)	484 (33.1)	480 (31.1)	0.262
TOTCHOL (mean (SD))	188.98 (41.91)	196.18 (40.34)	<0.001
DIABETES = 1 (%)	298 (19.1)	241 (14.8)	0.001
SYSBP_UT (mean (SD))	83.32 (59.99)	82.51 (57.33)	0.705
SYSBP_T (mean (SD))	44.13 (63.50)	42.06 (63.45)	0.373

We exclude some participants with missingness in cholesterol-related variables, BMI, and blood pressures. After omitting those missing values, we have 6% of missingness in `BPMEDS` and only 1 record missing in `DIABETES`. Then we applied multiple imputation technique to infer those missing values with 5 imputation datasets. This method is trying to handle with each missing entry by estimating multiple reliable values such as regression models, running analysis across those completed dataset, aggregating all previous analyses results and analyzing how far they spread out in terms of standard deviations and confidence intervals.

Transportability Analysis

We assume that outcome and covariate information is obtained from a simple random sample from the source population (the Framingham data, $S = 1$). Furthermore, covariate information is obtained from a simple random sample from the target population (the NHANES data in 2017, $S = 0$), and no outcome information is collected from the target population. We assume the following identifiability conditions: (1) independence of the outcome and the population S conditioning on covariates X; (2) the probability of being from the source population conditioning on covariates must be greater than 0 for every x with positive density in the target population. These tow fairly strong conditions will allow us to tailor the prediction model and assess its performance in the target population. Given 5 complete imputation datasets from the NHANES data and complete cases from the Framingham data, we will split each of them into training and test sets. To tailor the prediction model $g_{\hat{\beta}}(X)$ for use in the target population, we assume the model $g_{\beta}(X)$ is misspecified in most practical application cases. Then we estimate β using the weighted maximum likelihood estimator, which can be obtain from the inverse of the odds of being from the source population. Although the inverse-odds weights are not identifiable, we assume, up to unknown proportionally constant, they are equal to the inverse-odds weights in the training set $\frac{Pr(S=0)|X,D_{\text{train}}=1}{Pr(S=1)|X,D_{\text{train}}=1}$[1].

Specifically, we will use 80% of the sex-specific Framingham dataset as training set and 20% as test set, as well as sex-specific imputation datasets. Following the above inverse-odds weights in the training dataset, we firstly combine training set from the Framingham and from each of training imputed 2017-NHANES sets under women and men categories separately, and then fit the logistic model with respect to the population

S given covariates mentioned above to get the inverse odds of being from the source population. Since this estimator for the inverse-odds weights is only applicable in the source population, we use the `predict()` function on the training Framingham dataset and take the inverse of exponentiation of predicted outcomes. We tailored the prediction model by adding weights in the `glm()` function with respect to `CVD` and refit the model again to obtain the new estimated β coefficients. Given the tailored prediction model, we plugged into the Framingham test sets and set a threshold of 0.5 to cutoff the binary outcome 0 and 1. To get $\hat{\sigma}(X)$ of the inverse-odds weights in the test set $\frac{Pr(S=0)|X,D_{\text{test}}=1}{Pr(S=1)|X,D_{\text{test}}=1}$, we exponent results from the `predict()` function to the test Framingham data and calculate inverse. Given all those quantities, we estimate the Brier risk scores in the target population following the equation: $\hat{\phi}_{\beta} = \frac{\sum_{i=1}^{n} I(S_i=1,D_{\text{test},i}=1)\hat{\sigma}(X_i)(Y_i-g_{\hat{\beta}}(X_i))^2}{\sum_{i=1}^{n} I(S_i=0,D_{\text{test},i}=1)}$.

Table 3: Estimated Brier Scores in the Target Population

	Men	Women
Composite Fram and Imp 1 Test	0.168	0.041
Composite Fram and Imp 2 Test	0.189	0.044
Composite Fram and Imp 3 Test	0.196	0.034
Composite Fram and Imp 4 Test	0.190	0.053
Composite Fram and Imp 5 Test	0.199	0.041
Average Estimation for Brier Risk	0.188	0.041

Our results pertain to applications where the prediction model is built using the training data and is evaluated using the test data, and where the entire composite data set is split into a test and a training set that are used for model estimation and assessment[1]. Brier scores range between 0 and 1. A score of 0 represents perfect accuracy and a score of 1 represents perfect inaccuracy. Then, our results of Brier scores are closed to 0 by gender, which informed us a good performance of this tailored prediction model in this transportability analysis.

Illustration Using Simulated Data

Now we assume that individual level data is not available from the target population and only summary statistics with mean and standard deviation derived from the NHANES dataset are available. In the simulation study, we aim to investigate how estimation for the Brier score would be influenced in the simulated target population under different data generation processes for variables. The data generation method is varying in this section. In particular, we will focus on two ways. Firstly, we would like to get correlation insights from the source population (the Framingham study) by taking log transformation for each continuous variable. The second method is to determine the exact distribution followed by continuous variable in the Framingham data, and simulate it by

The shared parameter is the number of samples to generate (N = 3000) and significance level of 0.05. As for the first data generation method, we took log transformation for each continuous variable to ensure the normality and tested the correlation matrix. Given fixed mean and standard deviation derived from the NHANES data, we simulated from a multivariate normal distribution with defined mean vector and covariance matrix of the continuous variables with the consideration of standard deviation. Except for continuous variable, we also considered about potential correlation between categorical variables, such as use of medication, and continuous variables, such as systolic blood pressure. To get the sense of association, we fitted the logistic model to determine if some variables are highly associated with binary outcomes. The first binary variable we examined is `BPMEDS`. The p-value showed that systolic blood pressure plays an important role in predicting the odds of use of medications, thus, we predicted `BPMEDS` in the simulated target population given the logistic regression model with respect to `BPMEDS` given important continuous predictors in the source population. We followed the same procedure to simulate gender variable. Specifically, given all simulated continuous variables and one categorical variable, we fitted the logistic regression model again in the Framingham data and examine the significance of predictors. Based on the results, we applied the `predict()` function on log transformation of simulated dataset and set a threshold of 0.5. By continually fitting logistic regression models with respect to categorical covariates in the source population, we get primary ideas about potential association between categorical and continuous variables and simulate the proportion of binary covariates by estimated coefficients of significant predictors.

For the second data generation method, we particularly determined the distributions of continuous covariates in the source population, and simulated new individual-level data under certain parameters settings. We used `descdist()` function to compute descriptive parameters of an empirical distribution for non-censored data with bootstrapping method and provide a skewness-kurtosis plot. For example, the Cullen and Frey graph show us systolic blood pressure followed the gamma distribution in the source population, as the observation (blue) point is closed to the dashed line of theoretical gamma distribution. Then we used `fitdistr()` function to get the shape and rate parameters, and simulate the `SYSBP` under a gamma distribution with the shape parameter of 40.18 and the rate parameter of 0.29. Following the same step, we firstly determined theoretical distributions for each continuous variable based on Cullen and Frey graphs and used fixed mean and standard deviation values to simulate all continuous variables. In summary, `AGE` follows a normal distribution with mean of 52.41 and standard deviation of 12.60; systolic blood pressure `SYSBP` follows a gamma distribution with parameters mentioned above; `BMI` follows a log normal distribution with mean of 30.32 on the log scale and standard deviation of 7.33 on the log scale; `HDLC` follows a log normal distribution with mean of 52.89 on the log scale and standard deviation of 15.86 on the log scale; `TOTCHOL` follows a log normal distribution with mean of 192.66 on the log scale and standard deviation of 41.26 on the log scale. As for categorical variables, we follow the same procedure implemented in the first data generation method. Again, we fitted logistic regression models to understand association between specified categorical variables and other possible important covariates. The more variables we've simulated, the higher accuracy we can obtain for simulation of binary variables of interests.

Table 4: Estimated Brier Scores in the Simulated Target Population

	Men	Women
Method 1	0.1749	0.0334
Method 2	0.1346	0.0239

In our simulation design, our focus is on estimating Brier risk scores based on proposed methods in published work. Then we, at the same time, utilize these scores by gender to evaluate tailored model performance in the simulated target population. The Table 4 consists of four Brier scores under each data generation method by sex. We can find that all Brier scores are less than estimations obtained from the 2017 NHANES data, which represents our simulation, to some extent, mimic the true distributions of covariates in the sample NHANES data. However, it is trivial to observe the second data generation method performs better in Brier scores by gender than the first one as we extract information about distribution exactly from the source population. As we simulated cases are truly closed to distribution in the source population so that we obtained with a lower Brier score.

Limitations

The main limitation in our simulation study is the order of projection. For example, even though we found current smoking status and prevalence of diabetes play significant role in predicting participants' gender, due to the order of simulation process, we don't have information about cigarette smoking and diagnoses of diabetes prior to the simulation of gender, and are constrained to simulate correlated data at the current stage. Moreover, this simulation study doesn't account for other varying factors, such as number of simulations, number of samples to generate, and varying correlation matrices among covariates, which should be improved in the further studies.

Reference

1. Steingrimsson, Jon A., et al. “Transporting a prediction model for use in a new target population.” American Journal of Epidemiology 192.2 (2023): 296-304.

2. D’Agostino Sr, Ralph B., et al. “General cardiovascular risk profile for use in primary care: the Framingham Heart Study.” Circulation 117.6 (2008): 743-753.

3. Li, Bing, et al. “Estimating the area under the ROC curve when transporting a prediction model to a target population.” Biometrics 79.3 (2023): 2382-2393.