Assessing Predictive Model Transportability

Abstract

In this detailed study, we compare simulation-based and dataset-based methods to evaluate the transportability of the Framingham Heart Study's CVD risk prediction model to the NHANES population. Using logistic regression models, this analysis integrates data from both datasets, providing a thorough assessment of the model's performance in varying demographic environments. We employ Brier scores to meticulously assess the model's predictive accuracy. The results highlight significant variations in performance, emphasizing the critical need for model adaptation and validation across different population groups. This study not only illustrates the importance of tailored risk prediction models but also draws attention to potential limitations in data and methodological approaches in such cross-population studies.

Introduction

In the realm of predictive modeling, the primary objective is often to utilize the insights gained from a model built on one dataset and apply it to a different population of interest, known as the target population. For instance, healthcare systems frequently seek to deploy risk prediction models to identify individuals at high risk for specific health events, such as cardiovascular events, within their patient population. However, an inherent challenge arises when the data used to develop the prediction model, referred to as the source study data, do not perfectly align with the characteristics of the target population.

To illustrate this challenge, consider the Framingham ATP-III model, a widely-used tool for predicting the 10-year risk of cardiovascular events. This model was constructed using source data primarily comprised of participants from a single demographic, predominantly white individuals. However, research has demonstrated that this model's performance may not generalize effectively to diverse, multi-ethnic populations.

In response to these challenges, recent years have witnessed the development of various methods to assess and adapt prediction models for use in target populations. These methods aim to evaluate how well a model, initially developed on one population, performs when applied to a different population. This process often involves "transporting" measures of model performance from the source population to the target population (Steingrimsson et al., 2022).

The objective of this study is to apply these innovative methods for assessing the transportability of the Framingham Heart Study's cardiovascular disease (CVD) risk prediction model to a distinct and diverse population represented in the National Health and Nutrition Examination Survey (NHANES). Subsequently, we will utilize Brier scores, an effective metric for assessing predictive accuracy, to scrutinize and quantify the model's performance within the NHANES population.

Methods

Data Source and Preprocessing

Framingham Heart Stady Dataset

The Framingham Heart Study dataset, a widely-recognized resource for cardiovascular risk prediction, serves as the source dataset for our modeling efforts. In our modeling approach, we follow a methodology that mirrors the models presented in the paper "General Cardiovascular Risk Profile for Use in Primary Care."

The initial step involved selecting a subset of relevant variables from the Framingham dataset. The selected variables include:

- CVD status (Cardiovascular Disease)
- TIMECVD (time to cardiovascular disease event)

- SEX (gender)
- TOTCHOL (total cholesterol)
- AGE (age)
- SYSBP (systolic blood pressure)
- DIABP (diastolic blood pressure)
- CURSMOKE (current smoking status)
- DIABETES (diabetes status)
- BPMEDS (blood pressure medications)
- HDLC (high-density lipoprotein cholesterol)
- BMI (body mass index)

Adressing missing data, we excluded observations presenting missing values the dataset. This data preprocessing step is essential, as it helps ensure the quality of our analysis. It is crucial to note that this exclusion is based on the assumption that the missingness of data is not systematically related to the study outcomes.

Blood pressure measurements (SYSBP) were adjusted based on the use of BPMEDS (blood pressure medications). SYSBP_UT and SYSBP_T were derived, representing untreated and treated blood pressure, respectively. To assess cardiovascular risk within a 15-year timeframe, data filtering was performed. Observations with a TIMECVD (time to cardiovascular disease event) less than or equal to 15 years were retained, while those with longer follow-up periods were excluded from the analysis. The dataset was further divided into two subsets, one for males (framingham_df_men) and one for females (framingham_df_women), to allow for gender-specific risk modeling.

An initial exploratory analysis, stratified by SEX, offered insights into the demographic and clinical characteristics of the study cohort, helping to identify potential gender-based differences in risk factors (Table 1).

Characteristic	1, N = 1,094[1]	2, N = 1,445[1]	p-value[2]
CVD	360 (33%)	242 (17%)	<0.001
TOTCHOL	223 (198, 253)	243 (214, 272)	<0.001
AGE	60 (53, 67)	60 (53, 67)	0.13
SYSBP	136 (123, 151)	137 (122, 155)	0.6
DIABP	80 (75, 89)	80 (72, 87)	<0.001
CURSMOKE	425 (39%)	445 (31%)	<0.001
DIABETES	96 (8.8%)	95 (6.6%)	0.037
BPMEDS	123 (11%)	259 (18%)	<0.001
HDLC	42 (35, 50)	52 (42, 61)	<0.001
BMI	26.0 (24.0, 28.2)	24.8 (22.7, 27.7)	<0.001
SYSBP_UT	131 (118, 145)	127 (110, 144)	<0.001
SYSBP_T	0 (0, 0)	0 (0, 0)	<0.001

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

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

Table 1. Summary of Framingham Heart Study dataset Variables by Sex (1 – Men, 2 – Women)

NHANES Stady Dataset

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. This dataset is particularly valuable for

its comprehensive collection of data on various health-related variables, making it a vital resource for public health research and policy making.

NHANES combines interviews and physical examinations to gather detailed health information. The survey is unique in that it combines personal interviews with physical examinations, including laboratory tests conducted by highly trained medical personnel. This methodology provides a deep and well-rounded perspective on the health conditions and risk factors present in the U.S. population.

For the purpose of this study, the NHANES dataset serves as a complementary source to the Framingham Heart Study, facilitating a broader examination of cardiovascular risk factors in a more diverse population. The dataset was accessed using the 'nhanesA' library in R, which simplifies the process of extracting data from the extensive NHANES database.

Several key variables were extracted from different NHANES survey modules to create a dataset that mirrors the covariates used in the Framingham Heart Study. These variables included blood pressure measurements (SYSBP), demographic information (SEX and AGE), body mass index (BMI), smoking status (CURSMOKE), use of blood pressure medications (BPMEDS), total cholesterol (TOTCHOL), high-density lipoprotein cholesterol (HDLC), and diabetes status (DIABETES). Each variable was selected from the relevant survey module, and for consistency with the Framingham dataset, variables were renamed accordingly.

To ensure consistency and relevance for the subsequent analysis, some variables required additional processing. Notably, CURSMOKE was derived based on smoking-related variables (SMQ040 and SMQ020) to create a binary indicator for current smoking status, BPMEDS was created by considering multiple blood pressure-related questions (BPQ020, BPQ040A, and BPQ050A) to identify blood pressure medication usage, and DIABETES was derived from DIQ010 by categorizing individuals as having diabetes or not.

To create a comprehensive dataset, data from the different NHANES survey modules were combined. As a result, a single dataset ('df_2017') was created, containing information on all selected covariates from NHANES.

An initial exploratory examination, categorized by gender (SEX), provided valuable insights into the demographic and clinical attributes of the study cohort. Table 2 provides an overview of variable distributions and allows for comparisons between genders.

To ensure the validity of subsequent analyses, observations with missing values in any of the selected covariates were removed from the dataset, resulting in a dataset ('df_2017') with complete information.

The NHANES dataset was harmonized with the eligibility criteria outlined in the Framingham study. It involves filtering the NHANES data to include only individuals aged 30 to 59 and then segregating the dataset by gender.

Characteristic	1, N = 2,105[1]	2, N = 2,205[1]	p-value[2]
SYSBP	124 (114, 136)	120 (108, 134)	<0.001
AGE	52 (33, 65)	50 (33, 64)	0.025
BMI	28 (25, 33)	28 (24, 34)	0.5
HDLC	46 (39, 55)	56 (47, 67)	<0.001
CURSMOKE	429 (20%)	316 (14%)	<0.001
BPMEDS	627 (30%)	640 (29%)	0.6
TOTCHOL	179 (153, 209)	186 (162, 215)	<0.001
DIABETES	370 (18%)	271 (12%)	<0.001
SYSBP_UT	114 (0, 126)	108 (0, 120)	<0.001
SYSBP_T	0 (0, 118)	0 (0, 116)	>0.9

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

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

Table 1. Summary of NHANES dataset Variables by Sex (1 – Men, 2 – Women)

Model Development

For this project, separate logistic regression models were developed for men and women using the Framingham Heart Study dataset. The models predict the occurrence of cardiovascular disease (CVD) as a binary outcome based on several predictors: high-density lipoprotein cholesterol (HDLC), total cholesterol (TOTCHOL), age (AGE), systolic blood pressure (SYSBP), whether currently smoking (CURSMOKE), and diabetes status (DIABETES). The continuous predictors (HDLC, TOTCHOL, AGE, SYSBP) were log-transformed.

Simulation Design and Analysis

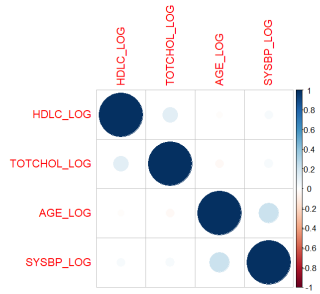
Aims

This simulation study aims to evaluate the transportability of a CVD risk prediction model from the Framingham Heart Study to the NHANES population, focusing on the model's predictive accuracy in a demographically and clinically distinct target population.

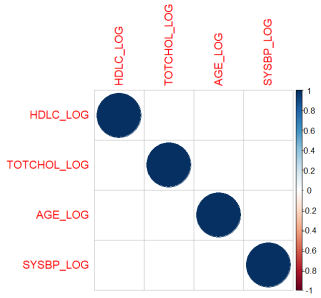
Data-generating Mechanisms

In the simulation study, data from the Framingham Heart Study and NHANES were merged into a combined dataset. While the Framingham dataset included CVD outcome data, this outcome data was not present in the NHANES dataset. Data-generating mechanism involves three distinct settings (Setting 1, Setting 2, and Setting 3) for data with a focus on understanding the impact of correlation between continuous variables.

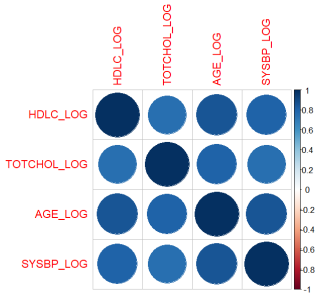
Setting 1 represents the correlation structure of the Framingham dataset. In Setting 2, the focus shifts to intentionally introducing no correlation between continuous variables. Setting 3 explores the effects of a high level of correlation between continuous variables. Pictures 1, 2, an3 illustrate correlations for three settings accordingly.



Picture 1. Correlations for Setting 1



Picture 2. Correlations for Setting 2



Picture 3. Correlations for Setting 3

Data is generated for each correlation setting, creating multivariate normal data with a specified mean and correlation matrix.

Estimands

This study utilizes Inverse Probability of Selection Weights (IPSW) to bridge the demographic and clinical differences between the Framingham Heart Study and the NHANES population. This statistical tool is instrumental in adapting the Framingham-derived cardiovascular disease (CVD) risk model for effective use in the NHANES context. The IPSW approach calculates the likelihood ratio of an individual belonging to the source (Framingham) versus the target (NHANES) population based on their covariates:

$$\frac{\Pr[S = 1 \mid X]}{\Pr[S = 0 \mid X]}$$

This ratio is crucial for mitigating biases introduced by population heterogeneity, ensuring the model's applicability and validity across diverse demographic and clinical settings.

Methods

The methodology hinges on logistic regression models to estimate the inverse odds of selection, a fundamental aspect of the transportability analysis. This approach facilitates the computation of the probabilities that an individual with specific characteristics belongs to the source population. The logistic models ensure alignment with the principles of transportability and validity in diverse populations. The estimations are further integrated into the overall model, enhancing its predictive power and relevance across different population groups.

$$\widehat{\psi} = \frac{\sum_{i=1}^{n} I(S_i = 1)\hat{o}(X_i)\left(Y_i - g(X_i)\right)^2}{\sum_{i=1}^{n} I(S_i = 0)}, \text{where } \hat{o}(X_i) = \frac{Pr[S = 0|X]}{Pr[S = 1|X]}$$

Performance Measures

The study employs Brier scores as the primary metric to evaluate the model's predictive accuracy. These scores, essentially representing the mean squared error of probabilistic forecasts, are pivotal in assessing the model's calibration and discrimination capabilities. A lower Brier score indicates a higher accuracy of the model in predicting actual outcomes, a critical aspect in the realm of CVD risk prediction. The application of Brier scores in various simulation settings offers profound insights into the model's adaptability and robustness, underlining its practical utility in diverse healthcare contexts.

Results

In this simulation study, the Brier score, a measure of predictive accuracy, was utilized to assess a CVD risk prediction model's performance when transported from the Framingham Heart Study to the NHANES population. The study implemented three distinct simulation settings to explore the effects of varying correlation structures among.

The results revealed that the Brier score for women in the NHANES population was 0.1200, and for men, it was 0.1164. In the Framingham population, the scores were slightly higher for men at 0.1919. Interestingly, applying different settings, Brier score increased, indicating a decrease in predictive accuracy, Specifically, in Setting 1, the Brier score for men was 0.2752 and for women was 0.3132. Settings 2 and 3 showed similar trends with scores for men being 0.2731 and 0.2808, and for women 0.3142 and 0.3131, respectively. The increase in Brier scores in settings with altered correlation structures suggests that the relationships among variables significantly affect model performance. This highlights the importance of understanding variable interactions in risk prediction.

Population	Brier Score Framingham	Brier Score NHANES	Setting 1 NHANES	Setting 2 NHANES	Setting 3 NHANES
Women	0.1160	0.1200	0.3132	0.3142	0.3131
Men	0.1919	0.1164	0.2752	0.2731	0.2808

Table 3. Brier Score for different datasets and settings for men and women

Additionally, in different simulation settings that explored varying correlation structures among continuous variables, the Brier scores increased for both genders. This increase highlights the potential impact of different correlation structures on the model's predictive accuracy. The findings underscore the importance of considering such variations and the need for careful model adjustment and validation when applying risk prediction models across diverse populations.

Setting	Population	Brier Score Mean	Brier Score SD
1	Men	0.2752	0.0032
	Women	0.3151	0.0019
2	Men	0.2749	0.0032
	Women	0.3149	0.0020
3	Men	0.2747	0.0035
	Women	0.3146	0.0022

Table 4. Brier Score for simulations with different settings for men and women

These outcomes suggest that while the Framingham model can be transported to the NHANES population, there is a notable variation in predictive accuracy, especially when considering different correlation structures

within the simulation settings. This highlights the need for careful adjustment and validation of risk prediction models when applying them to different populations.

Limitations

The study's limitations include potential biases from non-random sampling in the source data, potentially affecting model transportability. The lack of diverse demographic representation in the Framingham data may limit generalizability to the NHANES population. Additionally, the simulation approach, while informative, cannot fully replicate real-world conditions. There's also the assumption that missing data in both datasets do not systematically relate to outcomes, which may not hold true. These factors must be considered when interpreting the results and applying them to broader contexts.

References

1. D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation, 117(6), 743-753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579
2. Steingrimsson, J. A., Gatsonis, C., Li, B., & Dahabreh, I. J. (2022). Transporting a Prediction Model for Use in a New Target Population. American Journal of Epidemiology, 192(2). https://doi.org/10.1093/aje/kwac128
3. Morris, T. P., White, I. R., & Crowther, M. J. (2018). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11), 2074-2102. https://doi.org/10.1002/sim.8086