Evaluation of Cardiovascular Risk Prediction Models in NHANES Population

Liangkang Wang

2023-12-14

This report assesses the transportability of cardiovascular risk prediction models from the Framingham study to the NHANES cohort. Logistic regression models were built to predict CVD in men and women, evaluated by Brier scores and ROC curves within the Framingham data. These models were then applied to NHANES to examine their performance in a broader population.

Simulation studies mirrored the NHANES variable distributions and Framingham correlations to generate synthetic data, ensuring realistic gender-specific datasets for analysis. The results confirmed strong predictive accuracy in the original cohort and substantial accuracy in the NHANES dataset, particularly for women. The simulations illustrated that model transportability is contingent on the correlation structure between variables, emphasizing the need for precise data representation for model applicability.

In summary, the cardiovascular risk prediction models showed potential for generalization beyond their original cohort, indicating their viability for wider clinical and public health application. However, the effectiveness of these models across different populations requires careful consideration of data structures and demographic specifics.

Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality globally, and the development of robust risk prediction models is crucial for early intervention and prevention strategies. This project embarks on a critical evaluation of how well a cardiovascular risk prediction model, developed from the Framingham Heart Study, performs when applied to a different, more diverse population represented by the National Health and Nutrition Examination Survey (NHANES). The Framingham study, a pioneering research project started in 1948, significantly advanced our understanding of CVD risk factors. However, its findings have often been scrutinized for limited applicability beyond its largely homogeneous participant base,

predominantly of European descent. In contrast, NHANES offers a broader representation of the US population, encompassing a more diverse demographic.

The essence of this analysis lies in the concept of "transportability" – assessing whether a prediction model, developed in one specific setting or population (the Framingham study), retains its accuracy and reliability when applied to another distinct population (NHANES). This concept is pivotal in understanding and overcoming the challenges of model generalization in public health. By focusing on transportability, this study aims to bridge the gap between theoretical model development and practical, real-world application, ensuring that predictive tools for CVD risk are robust, reliable, and widely applicable across diverse populations.

Data Sources and Preparation

Framingham Heart Study Data

The Framingham Heart Study data set is an integral component of our analysis, serving as the source study data for the cardiovascular risk prediction model. This data set originates from the Framingham Heart Study, a pioneering long-term prospective study focused on the etiology of cardiovascular disease. Initiated in 1948 in Framingham, Massachusetts, the study initially enrolled 5,209 subjects and has since been instrumental in advancing our understanding of cardiovascular risk factors and their combined effects. The data set we are utilizing is a subset of this extensive study, encompassing laboratory, clinic, questionnaire, and adjudicated event data for 4,434 participants. These participants underwent examinations approximately every six years from 1956 to 1968, and each was followed for a total of 24 years. The data set includes detailed information on various parameters such as serum cholesterol levels, blood pressure, smoking history, body mass index (BMI), and diabetes status, along with outcomes like myocardial infarction, stroke, and death. It is a rich resource that provides comprehensive insights into cardiovascular health and disease progression, making it an ideal foundation for developing and evaluating prediction models for cardiovascular events Kornej et al. (2022).

NHANES Data

The National Health and Nutrition Examination Survey (NHANES) data is a crucial element of our study, offering a comprehensive look at various health and nutritional parameters of the U.S. population. NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States, and it is unique in that it combines interviews and physical examinations United States Department of, Human Services. Centers for Disease, and Prevention. National Center for Health (2012).

Variable	Male	Female	p-value
n	1094	1445	
CVD (mean (SD))	0.33 (0.47)	0.17 (0.37)	<0.001
SEX = 2 (%)	0 (0.0)	1445 (100.0)	<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
DIABP (mean (SD))	81.99 (11.31)	80.34 (11.06)	<0.001
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
SYSBP_UT (mean (SD))	121.04 (46.69)	111.49 (55.89)	<0.001
SYSBP_T (mean (SD))	17.90 (50.93)	28.45 (61.53)	<0.001

Note:

Table 1: Framingham Summary

Variable	Male	Female	p-value
n	2105	2205	
SEQN (mean (SD))	98306.09 (2714.29)	98285.62 (2686.95)	0.804
SYSBP (mean (SD))	126.44 (16.83)	123.70 (20.36)	<0.001
SEX = 2 (%)	0 (0.0)	2205 (100.0)	<0.001
AGE (mean (SD))	50.15 (18.83)	48.90 (18.57)	0.029
BMI (mean (SD))	29.19 (6.25)	29.84 (7.96)	0.003
HDLC (mean (SD))	48.11 (13.59)	58.10 (15.68)	<0.001
CURSMOKE = 1 (%)	429 (20.4)	316 (14.3)	<0.001
BPMEDS = 1 (%)	627 (29.8)	640 (29.0)	0.607
TOTCHOL (mean (SD))	183.10 (41.65)	190.51 (41.20)	<0.001
DIABETES = 1 (%)	370 (17.6)	271 (12.3)	<0.001
SYSBP_UT (mean (SD))	86.46 (57.73)	83.64 (55.42)	0.101
SYSBP_T (mean (SD))	39.98 (62.19)	40.07 (63.62)	0.964

Note:

Table 2: NHANES Summary

Matching Variables Descriptions

In our analysis, we have utilized several key variables from the Framingham Heart Study and NHANES data sets, each serving a distinct role in understanding cardiovascular risk factors. Here are the descriptions of these variables:

- `CVD`: Represents the occurrence of cardiovascular disease. It is a binary variable, where 1 indicates the presence of cardiovascular disease and 0 indicates its absence.
- `SEX`: Participant's sex, where 1 denotes male and 2 denotes female.
- `TOTCHOL`: Serum Total Cholesterol measured in mg/dL.
- `AGE`: Age of the participant at the time of the examination, measured in years.
- `SYSBP`: Systolic Blood Pressure, measured in mmHg. It represents the pressure in blood vessels when the heart beats.
- `DIABP`: Diastolic Blood Pressure, measured in mmHg. It represents the pressure in blood vessels between heartbeats.
- `CURSMOKE`: Indicates current smoking status. 1 for current smokers, 0 for non-smokers.
- `DIABETES`: Indicates whether the participant is diabetic. 1 for diabetic, 0 for non-diabetic.
- `BPMEDS`: Indicates the use of anti-hypertensive medication. 1 for current use, 0 for not used.
- `HDLC`: High-Density Lipoprotein Cholesterol, measured in mg/dL. Available only for the third examination period.
- `BMI`: Body Mass Index, calculated as weight in kilograms divided by the square of height in meters.
- `SYSBP_UT`: Systolic Blood Pressure for participants not on anti-hypertensive medication (BPMEDS = 0).
- `SYSBP_T`: Systolic Blood Pressure for participants on anti-hypertensive medication (BPMEDS = 1).

We can find that `CVD` and `DIABP` only occur in `framingham` data set, and don't occur in the target population data set. `CVD` is the outcome variable, and `DIABP` needs to be removed when merging our data set.

In our model fitting, we only use `TOTCHOL`, `AGE`, `CURSMOKE`, `DIABETES`, `HDLC`, `SYSBP_UT`, `SYSBP_T` in our model fitting.

Data Integration and Preprocessing

In this report, our primary focus is on the transportability analysis across two distinct populations: the original population, where the model is fitted, and the target population. To maintain consistency and clarity in our analysis, we have chosen to exclude all observations with missing values from both datasets. This approach streamlines our study, as missing data and the potential selection bias arising from handling such data are not central to our investigation.

We split the Framingham dataset into a trainning set (70%) and a test set (30%). And we put all the NHANES data set into the test set.

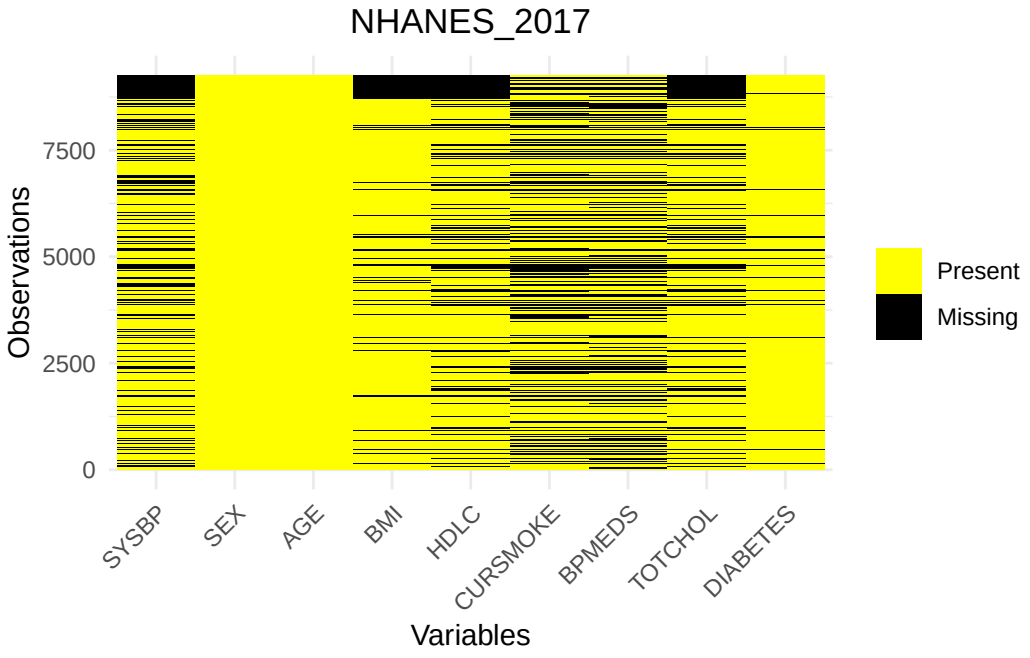

Figure 1: Missing heatmap of NHANES 2017

- Feature selection and engineering.

Methodology

Model Building

In our analysis, two distinct logistic regression models were developed to predict cardiovascular disease (CVD) incidence, separately for men and women. Both models use a binomial distribution, appropriate for binary outcome variables like CVD presence or absence.

For the male population, the model (mod_men) was built using the training dataset specifically filtered for observations from the Framingham study. The predictors in this model include the natural logarithm of high-density lipoprotein cholesterol (HDLC), total cholesterol (TOTCHOL), age (AGE), systolic blood pressure under treatment (SYSBP_UT), systolic blood pressure without treatment (SYSBP_T), current smoking status (CURSMOKE), and diabetes status (DIABETES). The inclusion of logarithmic transformations for continuous variables like HDLC, TOTCHOL, AGE, SYSBP_UT, and SYSBP_T suggests a non-linear relationship with the probability of developing CVD.

Similarly, the model for the female population (mod_women) follows the same structure and variable selection, ensuring that the analysis is consistent across genders.

Model Evaluation Metrics

In our methodology, to evaluate the performance of the two logistic regression models developed for predicting cardiovascular disease in men and women, we employ the Brier score and the Area Under the Receiver Operating Characteristic Curve (AUC/ROC). The Brier score is a metric for assessing the accuracy of probabilistic predictions. It measures the mean squared difference between the predicted probability assigned to the possible outcomes and the actual outcome. The AUC/ROC is a measure of a model's ability to distinguish between the two classes (presence or absence of cardiovascular disease in this case). It evaluates the model's performance across all classification thresholds, providing an aggregate measure of performance across all possible classification thresholds.

We evaluate our models performance on the testing set of framingham data set.

Transportability Analysis

In our study, we implement specific techniques to estimate the performance of our model when applied to the NHANES population. A key aspect of this analysis is the use of a tailored statistical formula, which allows us to quantify model performance accurately in the context of transportability.

The formula we use is as follows:

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

where $\hat{o}(X_i) = \frac{Pr(S=0|X_i, D_{test,i}=1)}{Pr(S=1|X_i, D_{test,i}=1)}$

In this equation, $\hat{\psi}_{\hat{\beta}}$ represents the estimated model performance in the NHANES population. The numerator sums the squared differences between the observed outcomes Y_i and the predicted probabilities $g_{\hat{\beta}}(X_i)$, weighted by $\hat{o}(X_i)$, for all individuals in the test dataset who are part of the NHANES study (indicated by $D_{test,i} = 1$). The denominator normalizes this sum, focusing on individuals not part of the Framingham study (indicated by $S_i = 0$) but included in the test data. The weight $\hat{o}(X_i)$ is the odds ratio of being in the NHANES population versus the Framingham population, conditional on the covariates X_i and being in the test data.

This formula is pivotal in our analysis as it adjusts for the differences in the distribution of covariates between the Framingham and NHANES populations, providing a more accurate assessment of how well the model performs when 'transported' from one population to another.

Simulation

The visual inspection of continuous variable distributions within the training dataset from the Framingham study, illustrated in Figure 3, suggests that they approximate a censored normal distribution, truncated at values greater than 0. Notably, the variables SYSBP_T and SYSBP_UT contain numerous zero values, influenced by the BPMED variable. In our simulation strategy, we plan to first simulate the SYSBP and BPMED variables, from which we will derive SYSBP_T and SYSBP_UT. The inter-variable relationships are evidenced by a correlation map and quantified in a covariance matrix.

For the simulation of the dataset, we utilize the means and standard deviations from the summary statistics of the NHANES dataset. This procedure marries the correlation structure of the Framingham dataset variables with the distributional parameters from NHANES, enabling us to craft a simulated dataset that reflects both sources.

The generation of categorical variables, namely `BPMED`, `SMOKE`, and `DIABETES`, is based on a binary distribution that mirrors the proportions observed in the NHANES dataset. These simulations are conducted separately for male and female datasets to account for gender-specific variations.

We have determined that the correlation matrix obtained from the NHANES data is the most representative of the true population correlations, while the matrix from the Framingham data is considered the least representative. By interpolating between these two matrices, we create four distinct correlation matrices that represent a spectrum from the least to the most representative. This gradient allows us to explore the effects of varying correlation structures on the transportability of our original model, providing insights into how such differences may impact model performance across different populations.

Results

Model Performance in the Framingham Study

Our cardiovascular disease prediction models demonstrate commendable predictive accuracy, as reflected by the Brier scores, which are 0.1912 for men and an even lower 0.1129 for women within the Framingham study group. These scores suggest that the predictions are closely aligned with the observed outcomes, particularly for women, indicating a high level of reliability in the model's probabilistic forecasts.

Expanding on this analysis, the ROC curves reveal an AUC of 0.725 for men and 0.778 for women, indicating that both models are adept at distinguishing between patients with and without cardiovascular disease. The higher AUC for women suggests that the model is especially effective for this group, managing to achieve a greater true positive rate for any given false positive rate.

```
[1] "Brier Score in Framingham for men is 0.191227391868493"
```

```
[1] "Brier Score in Framingham for women is 0.112948078539371"
```

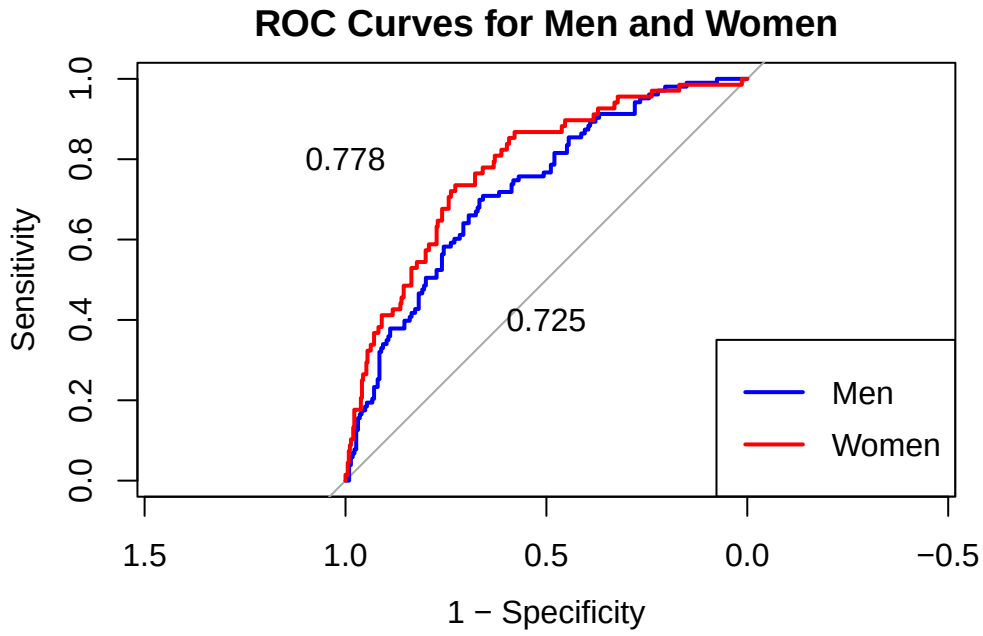

Figure 2: ROC Curves for Men and Women in Framingham

Transported Model Performance in NHANES

When assessing the transported model performance within the NHANES data, we observe promising results. The Transportation Brier scores, which measure the accuracy of the model's probabilistic predictions, are low for both men and women. Specifically, the Brier score for men is 0.0761, and for women, it is even lower at 0.0469.

These scores indicate a high level of accuracy in the transported models' predictions, with the model for women showcasing exceptional precision. A lower Brier score is desirable as it reflects closer agreement between the predicted probabilities and the actual outcomes. The considerably low Brier scores in the NHANES data compared to the Framingham data suggest that the models are not only generalizable but also potentially more accurate in this new population.

The difference in performance between genders is also noteworthy, with the model for women outperforming that for men, as indicated by the lower Brier score. This suggests that the factors influencing cardiovascular disease predictions may be better captured by the model in women or that the distribution of these factors in the NHANES population aligns more closely with the model's parameters for women.

```
[1] "Transportation Brier Score for men is 0.0761722727935116"
```

```
[1] "Transportation Brier Score for women is 0.046999634859797"
```

Simulation Study Results

In Figure 3 , we observed that the variable distributions for AGE, HDLC, SYSBP, SYSBP_T, SYSBP_UT, and TOTCHOL approximate normal distributions, albeit with a positive skew for SYSBP-related measures due to a clustering of values at zero. This skewness is a result of the censored nature of the data where the SYSBP_T and SYSBP_UT variables are derived from the BPMED variable, leading to a preponderance of zeros in their distribution as seen in Figure 3. The correlation map in Figure 4 showcases the relationship between these continuous variables, which is crucial for understanding the interdependencies within the cardiovascular risk factors.

The Brier scores for the simulation datasets, as shown in Table 3, reveal distinct differences between the four correlation settings, particularly between genders. For males, the scores range from 0.133 to 0.293, while for females, they are significantly lower, ranging from approximately 0.100 to 0.565. These differences in Brier scores can be attributed to the varying correlation structures used in the simulation. The first setting likely represents a baseline or control scenario, while subsequent settings incorporate varying degrees of correlation strength, reflecting the nuanced interplay between risk factors.

The most notable discrepancy is observed in the second correlation setting for females, where the Brier score is substantially higher than in other settings. This suggests that the specific correlation matrix used in this setting may not align well with the actual inter-variable relationships for the female cohort, leading to a less accurate simulation of the data and hence a higher Brier score. Conversely, the lower scores in other settings for females suggest a better capture of the underlying data structure, resulting in more accurate simulated datasets and, subsequently, better model performance.

The variability of the Brier scores across different correlation settings underscores the sensitivity of model performance to the assumed underlying data structure. It highlights the importance of accurately capturing the true correlations within the data to ensure the reliability and transportability of the cardiovascular risk prediction model.

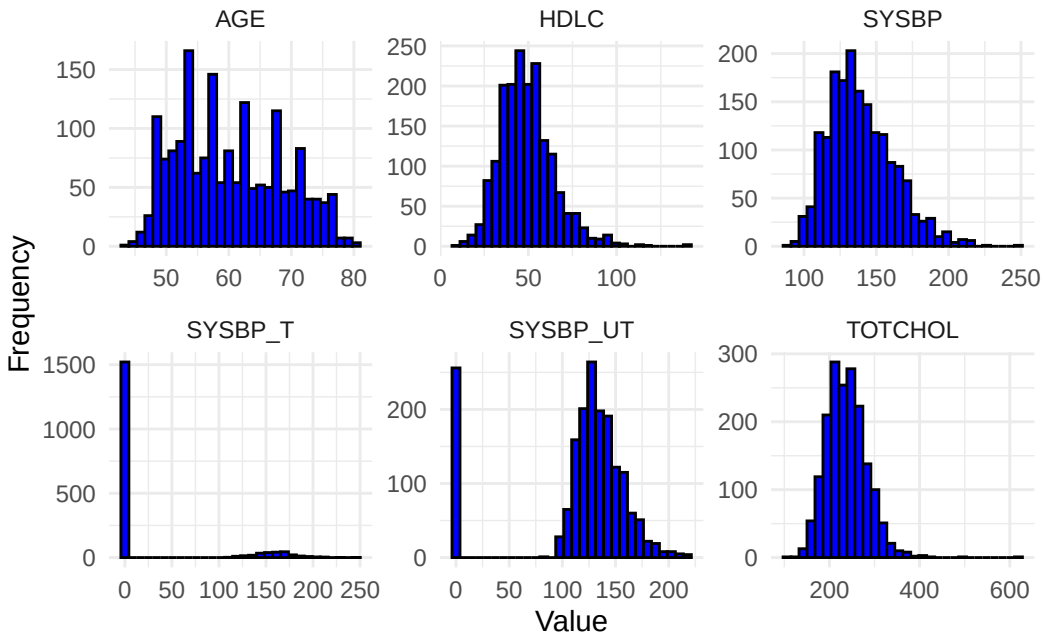

Figure 3: Distribution of Coninuous Variables

	Male	Female
Correlation setting: 1	0.2929899	0.0998615
Correlation setting: 2	0.1336183	0.5659993
Correlation setting: 3	0.2012964	0.1121319
Correlation setting: 4	0.2083992	0.1424570

Note:

Table 3: Brier Scores of simulation data sets

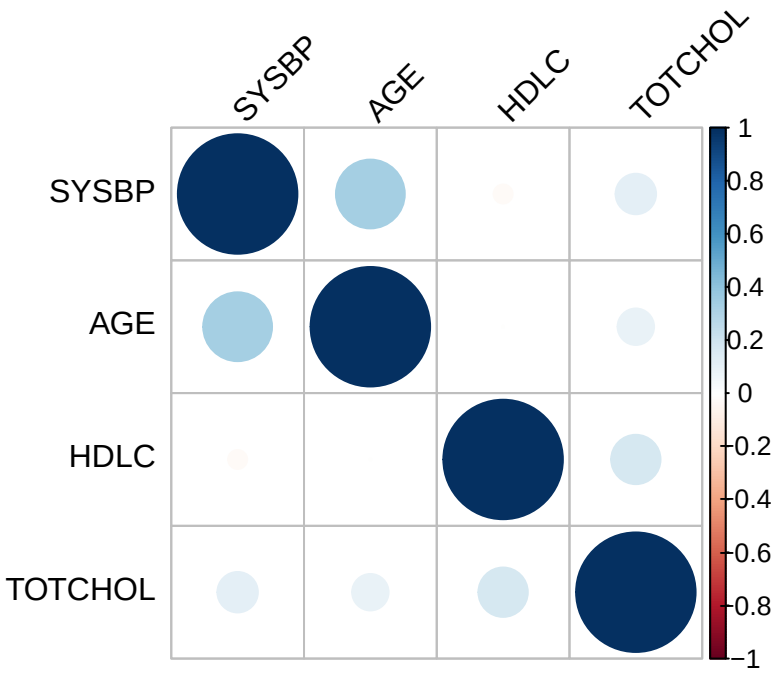

Figure 4: Correlation Maps of Coninuous Variables

Conclusion

In conclusion, our transportability analysis of cardiovascular risk prediction models provides valuable insights into the applicability of these models across different populations. The main findings from the Framingham and NHANES datasets underscore the robustness of our models, with the Brier scores indicating a high level of predictive accuracy in both the original and target populations.

The transportability analysis reveals that the models maintain their predictive power when applied to the NHANES cohort, with the Brier scores in the NHANES population being notably lower than those in the Framingham population, particularly for women. This indicates that the risk factors and their interactions captured by the models are indeed generalizable and relevant to a broader population.

The simulation studies, informed by the distribution and correlation of risk factors from both the Framingham and NHANES studies, highlight the importance of accurately capturing the underlying structure of the data. The variability in Brier scores across different simulated correlation structures emphasizes the need for precise modeling of inter-variable relationships to ensure the reliability and accuracy of the predictions when models are transported to different populations.

References

Kornej, J., D. Ko, H. Lin, J. M. Murabito, E. J. Benjamin, L. Trinquart, and S. R. Preis. 2022. "The Association Between Social Network Index, Atrial Fibrillation, and Mortality in the Framingham Heart Study." Journal Article. *Sci Rep* 12 (1): 3958. https://doi.org/10.1038/s41598-022-07850-9.

United States Department of, Health, Control Human Services. Centers for Disease, and Statistics Prevention. National Center for Health. 2012. "National Health and Nutrition Examination Survey (NHANES), 1999-2000." Dataset. Inter-university Consortium for Political; Social Research [distributor]. https://doi.org/10.3886/ICPSR25501.v4.

Appendices

```
library(riskCommunicator)
library(tidyverse)
library(tableone)
library(nhanesA)
library(knitr)
library(pROC)
library(kableExtra)
library(corrplot)
library(MASS)
source("Utils.R")
# library(Metrics)

opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(warning = FALSE, message = FALSE, echo = FALSE,
    fig.align = "center")
missing_heatmap <- function(data, title, color_pre = "yellow",
    color_miss = "black") {
    missing_values <- is.na(data)

    # Melt the matrix for use with ggplot
    missing_melted <- reshape2::melt(missing_values, id.vars = rownames(missing_values))

    # Create the heatmap
    g <- ggplot2::ggplot(missing_melted, aes(x = Var2, y = Var1)) +
        geom_tile(aes(fill = value)) + scale_fill_manual(name = "",
        labels = c("Present", "Missing"), values = c(color_pre,
            color_miss)) + theme_minimal() + theme(axis.text.x = element_text(angle = 45,
        hjust = 1)) + labs(x = "Variables", y = "Observations",
        title = title) + theme(plot.title = element_text(hjust = 0.5))
    return(g)
}
library(tidyverse)

data("framingham")

# The Framingham data has been used to create models for
# cardiovascular risk.  The variable selection and model
# below are designed to mimic the models used in the paper
```

```r
# General Cardiovascular Risk Profile for Use in Primary
# Care This paper is available (cvd_risk_profile.pdf) on
# Canvas.

framingham_df <- framingham %>%
    dplyr::select(c(CVD, TIMECVD, SEX, TOTCHOL, AGE, SYSBP, DIABP,
        CURSMOKE, DIABETES, BPMEDS, HDLC, BMI))

framingham_df <- na.omit(framingham_df)

# CreateTableOne(data=framingham_df, strata = c('SEX'))

# Get blood pressure based on whether or not on BPMEDS
framingham_df$SYSBP_UT <- ifelse(framingham_df$BPMEDS == 0, framingham_df$SYSBP,
    0)
framingham_df$SYSBP_T <- ifelse(framingham_df$BPMEDS == 1, framingham_df$SYSBP,
    0)

# Looking at risk within 15 years - remove censored data

# dim(framingham_df)
framingham_df <- framingham_df %>%
    filter(!(CVD == 0 & TIMECVD <= 365 * 15)) %>%
    dplyr::select(-c(TIMECVD))
# dim(framingham_df)

framingham_df$SEX <- as.factor(framingham_df$SEX)
framingham_df$DIABETES <- as.factor(framingham_df$DIABETES)
framingham_df$CURSMOKE <- as.factor(framingham_df$CURSMOKE)
framingham_df$BPMEDS <- as.factor(framingham_df$BPMEDS)

a <- CreateTableOne(data = framingham_df, strata = c("SEX"))
table_matrix <- print(a, printToggle = FALSE, noSpaces = TRUE)[,
    -4]
colnames(table_matrix)[1] <- "Male"
colnames(table_matrix)[2] <- "Female"

# Use kable to render the table kable(table_matrix, format
# = 'latex', booktabs = TRUE, col.names = c('Variable',
# 'Male', 'Female', 'p-value'))
```

```r
# kable(table_matrix, format = 'latex', booktabs = TRUE,
# col.names = c('Variable', 'Male', 'Female', 'p-value'))
# %>% kable_styling(latex_options = c('hold_position')) %>%
# add_footnote(caption = 'framingham summary',
# footnote_as_chunk = TRUE, threeparttable = TRUE)

# Load the kableExtra package library(kableExtra)

# Create a caption as a footnote
caption_text <- "Table 1: Framingham Summary"

# Render the table with a caption as a footnote using
# kableExtra
kable(table_matrix, format = "latex", booktabs = TRUE, col.names = c("Variable",
    "Male", "Female", "p-value")) %>%
    kable_styling(latex_options = c("hold_position")) %>%
    footnote(general = caption_text)

# The NHANES data here finds the same covariates among this
# national survey data
library(nhanesA)

# blood pressure, demographic, bmi, smoking, and
# hypertension info
bpx_2017 <- nhanes("BPX_J") %>%
    dplyr::select(SEQN, BPXSY1) %>%
    rename(SYSBP = BPXSY1)
demo_2017 <- nhanes("DEMO_J") %>%
    dplyr::select(SEQN, RIAGENDR, RIDAGEYR) %>%
    rename(SEX = RIAGENDR, AGE = RIDAGEYR)
bmx_2017 <- nhanes("BMX_J") %>%
    dplyr::select(SEQN, BMXBMI) %>%
    rename(BMI = BMXBMI)
smq_2017 <- nhanes("SMQ_J") %>%
    mutate(CURSMOKE = case_when(SMQ040 %in% c(1, 2) ~ 1, SMQ040 ==
        3 ~ 0, SMQ020 == 2 ~ 0)) %>%
    dplyr::select(SEQN, CURSMOKE)
bpq_2017 <- nhanes("BPQ_J") %>%
    mutate(BPMEDS = case_when(BPQ020 == 2 ~ 0, BPQ040A == 2 ~
        0, BPQ050A == 1 ~ 1, TRUE ~ NA)) %>%
    dplyr::select(SEQN, BPMEDS)
```

```r
tchol_2017 <- nhanes("TCHOL_J") %>%
    dplyr::select(SEQN, LBXTC) %>%
    rename(TOTCHOL = LBXTC)
hdl_2017 <- nhanes("HDL_J") %>%
    dplyr::select(SEQN, LBDHDD) %>%
    rename(HDLC = LBDHDD)
diq_2017 <- nhanes("DIQ_J") %>%
    mutate(DIABETES = case_when(DIQ010 == 1 ~ 1, DIQ010 %in%
        c(2, 3) ~ 0, TRUE ~ NA)) %>%
    dplyr::select(SEQN, DIABETES)

# Join data from different tables
df_2017 <- bpx_2017 %>%
    full_join(demo_2017, by = "SEQN") %>%
    full_join(bmx_2017, by = "SEQN") %>%
    full_join(hdl_2017, by = "SEQN") %>%
    full_join(smq_2017, by = "SEQN") %>%
    full_join(bpq_2017, by = "SEQN") %>%
    full_join(tchol_2017, by = "SEQN") %>%
    full_join(diq_2017, by = "SEQN")


# Get blood pressure based on whether or not on BPMEDS
df_2017$SYSBP_UT <- ifelse(df_2017$BPMEDS == 0, df_2017$SYSBP,
    0)
df_2017$SYSBP_T <- ifelse(df_2017$BPMEDS == 1, df_2017$SYSBP,
    0)

df_2017$SEX <- as.factor(df_2017$SEX)
df_2017$DIABETES <- as.factor(df_2017$DIABETES)
df_2017$CURSMOKE <- as.factor(df_2017$CURSMOKE)
df_2017$BPMEDS <- as.factor(df_2017$BPMEDS)

df_2017.omit <- na.omit(df_2017)

b <- CreateTableOne(data = df_2017.omit, strata = c("SEX"))
table_matrix <- print(b, printToggle = FALSE, noSpaces = TRUE)[,
    -4]
colnames(table_matrix)[1] <- "Male"
colnames(table_matrix)[2] <- "Female"
```

```r
# Use kable to render the table kable(table_matrix, format
# = 'latex', booktabs = TRUE, col.names = c('Variable',
# 'Male', 'Female', 'p-value')) knitr::kable(table_matrix,
# format = 'latex', booktabs = TRUE)

# kable(table_matrix, format = 'latex', booktabs = TRUE,
# col.names = c('Variable', 'Male', 'Female', 'p-value'))
# %>% kable_styling(latex_options = c('hold_position')) %>%
# add_footnote(caption = 'NHANES summary',
# footnote_as_chunk = TRUE, threeparttable = TRUE)


# Create a caption as a footnote
caption_text <- "Table 2: NHANES Summary"

# Render the table with a caption as a footnote using
# kableExtra
kable(table_matrix, format = "latex", booktabs = TRUE, col.names = c("Variable",
    "Male", "Female", "p-value")) %>%
    kable_styling(latex_options = c("hold_position")) %>%
    footnote(general = caption_text)
# missing_heatmap(data = framingham_df, title =
# 'framingham','yellow','black')
missing_heatmap(data = df_2017[, -c(1, 11, 12)], title = "NHANES_2017",
    "yellow", "black")

df_2017 <- na.omit(df_2017)
# Remove DIABP from framingham_df

framingham_df <- framingham_df %>%
    dplyr::select(-c(DIABP)) %>%
    mutate(SEQN = 0)
df_2017 <- df_2017 %>%
    mutate(CVD = 0)

framingham_df <- framingham_df %>%
    mutate(IS.framingham = 1)
df_2017 <- df_2017 %>%
    mutate(IS.framingham = 0)

#### Combined data set
```

```r
framingham_df_ordered <- framingham_df[names(df_2017)]



#### Split train and test set.

set.seed(123)

total_rows <- nrow(framingham_df_ordered)  # Get the total number of rows in the dataframe
train_size <- round(total_rows * 0.7)

train_indices <- sample(1:total_rows, train_size)

train_set <- framingham_df_ordered[train_indices, ]
test_set <- rbind(framingham_df_ordered[-train_indices, ], df_2017)

train_set_men <- train_set[train_set$SEX == 1, ]
train_set_women <- train_set[train_set$SEX == 2, ]

test_set_men <- test_set[test_set$SEX == 1, ]
test_set_women <- test_set[test_set$SEX == 2, ]


# Fit models with log transforms for all continuous
# variables.  We used the train dataset, which is a part of
# framingham dataset.  Also, Make two models for men and
# women separately.

mod_men <- glm(CVD ~ log(HDLC) + log(TOTCHOL) + log(AGE) + log(SYSBP_UT +
    1) + log(SYSBP_T + 1) + CURSMOKE + DIABETES, data = train_set_men[train_set_men$IS.fra
    1, ], family = "binomial")


mod_women <- glm(CVD ~ log(HDLC) + log(TOTCHOL) + log(AGE) +
    log(SYSBP_UT + 1) + log(SYSBP_T + 1) + CURSMOKE + DIABETES,
    data = train_set_women[train_set_women$IS.framingham == 1,
        ], family = "binomial")

# For Men
Y.fram.test.pred.men <- predict(mod_men, test_set_men[test_set_men$IS.framingham ==
    1, ], type = "response")
```

```r
Y.fram.test.true.men <- test_set_men$CVD[test_set_men$IS.framingham ==
    1]

# For Women
Y.fram.test.pred.women <- predict(mod_women, test_set_women[test_set_women$IS.framingham =
    1, ], type = "response")
Y.fram.test.true.women <- test_set_women$CVD[test_set_women$IS.framingham ==
    1]

# ROC for men in framingham

roc_men <- pROC::roc(Y.fram.test.true.men, Y.fram.test.pred.men)
auc_men <- pROC::auc(roc_men)

brier_score_men <- mean((Y.fram.test.pred.men - Y.fram.test.true.men)^2)


# ROC for women in framingham
roc_women <- pROC::roc(Y.fram.test.true.women, Y.fram.test.pred.women)
auc_women <- pROC::auc(roc_women)

brier_score_women <- mean((Y.fram.test.pred.women - Y.fram.test.true.women)^2)


paste0("Brier Score in Framingham for men is ", brier_score_men)
paste0("Brier Score in Framingham for women is ", brier_score_women)

# brier_score_men brier_score_women Assuming you have
# already calculated roc_men and roc_women using the roc
# function as shown in previous examples

# Plot the ROC curve for men
plot(roc_men, col = "blue", main = "ROC Curves for Men and Women",
    xlab = "1 - Specificity", ylab = "Sensitivity")

# Add the ROC curve for women to the same plot
lines(roc_women, col = "red")

# Add a legend to the plot
legend("bottomright", legend = c("Men", "Women"), col = c("blue",
    "red"), lwd = 2)
```

```r
text(0.5, 0.4, round(auc_men, digits = 3))
text(1, 0.8, round(auc_women, digits = 3))
# # Print out the AUC and Brier scores print(paste('AUC for
# men:', auc_men)) print(paste('Brier score for men:',
# brier_score_men)) print(paste('AUC for women:',
# auc_women)) print(paste('Brier score for women:',
# brier_score_women))

Tran_Brier <- function(test_set, Y.fram.test.true, Y.fram.test.pred) {

    Ps_X <- glm(IS.framingham ~ log(HDLC) + log(TOTCHOL) + log(AGE) +
        log(SYSBP_UT + 1) + log(SYSBP_T + 1) + CURSMOKE + DIABETES,
        data = test_set, family = "binomial")

    O_X <- (1 - Ps_X$fitted.values)/Ps_X$fitted.values
    O_X.fram <- O_X[test_set$IS.framingham == 1]

    phi_beta <- sum(O_X.fram * (Y.fram.test.true - Y.fram.test.pred)^2)/sum(test_set$IS.fr
        0)

    phi_beta
}


# prediction for Pr(S|X_i,D_{test,i}=1) with glm X
# CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+log(SYSBP_T+1)+CURSMOKE+DIABETES

# for men

phi_beta_men <- Tran_Brier(test_set_men, Y.fram.test.true.men,
    Y.fram.test.pred.men)

paste0("Transportation Brier Score for men is ", phi_beta_men)

# for women
phi_beta_women <- Tran_Brier(test_set_women, Y.fram.test.true.women,
    Y.fram.test.pred.women)

paste0("Transportation Brier Score for women is ", phi_beta_women)

continuous_vars <- train_set[, c("SYSBP_T", "SYSBP_UT", "SYSBP",
```

```
    "AGE", "HDLC", "TOTCHOL")]
continuous_vars_nhanes <- test_set[test_set$IS.framingham ==
    0, c("SYSBP_T", "SYSBP_UT", "SYSBP", "AGE", "HDLC", "TOTCHOL")]

long_data <- continuous_vars %>%
    gather(key = "Variable", value = "Value")

ggplot(long_data, aes(x = Value)) + geom_histogram(bins = 30,
    fill = "blue", color = "black") + facet_wrap(~Variable, scales = "free",
    nrow = 2) + theme_minimal() + labs(x = "Value", y = "Frequency")

cor_matrix <- cor(continuous_vars[], use = "complete.obs")
cor_matrix_NHANES <- cor(continuous_vars_nhanes, use = "complete.obs")

cov_matrix <- cov(continuous_vars, use = "complete.obs")

corrplot(cor_matrix[3:6, 3:6], method = "circle", tl.col = "black",
    tl.srt = 45)
cor_matrix_sim_1 <- cor_matrix[3:6, 3:6]
cor_matrix_NHANES_sim <- cor_matrix_NHANES[3:6, 3:6]

cor_matrix_sim_2 <- cor_matrix_sim_1 * 0.3 + cor_matrix_NHANES_sim *
    0.7
cor_matrix_sim_3 <- cor_matrix_sim_1 * 0.5 + cor_matrix_NHANES_sim *
    0.5
cor_matrix_sim_4 <- cor_matrix_sim_1 * 0.7 + cor_matrix_NHANES_sim *
    0.3


nrow_men = 2105
nrow_women = 2205

P_cate_men = c(429, 370, 627)/2105
P_cate_women = c(316, 271, 640)/2205

# SYSBP,AGE,HDLC,TOTCHOL
mean_vec_men = c(126.44, 50.15, 48.11, 183.1)
mean_vec_women = c(123.7, 48.9, 58.1, 190.51)

std_vec_men = c(16.83, 18.83, 13.59, 41.65)
std_vec_women = c(20.36, 18.57, 15.68, 41.2)
```

```r
### Get covariance matrix from Framingham training set.
### Substitute the variance with what we got from
### NHanes_table

sim_data <- function(cor_matrix, std_vec, mean_vec, p_vec, nrow,
    test_set) {

    set.seed(123)

    cov_matrix <- cor_matrix * (std_vec %*% t(std_vec))

    simulated_data <- MASS::mvrnorm(nrow, mu = mean_vec, Sigma = cov_matrix)

    simulated_CURSMOKE <- rbinom(n = nrow, size = 1, prob = p_vec[1])
    simulated_DIABETES <- rbinom(n = nrow, size = 1, prob = p_vec[2])
    simulated_BPMEDS <- rbinom(n = nrow, size = 1, prob = p_vec[3])

    #
    simulated_data <- data.frame(simulated_data)

    simulated_data$CURSMOKE <- simulated_CURSMOKE
    simulated_data$DIABETES <- simulated_DIABETES
    simulated_data$BPMEDS <- simulated_BPMEDS

    #
    colnames(simulated_data) <- c("SYSBP", "AGE", "HDLC", "TOTCHOL",
        "CURSMOKE", "DIABETES", "BPMEDS")

    simulated_data$SYSBP_UT <- ifelse(simulated_data$BPMEDS ==
        0, simulated_data$SYSBP, 0)
    simulated_data$SYSBP_T <- ifelse(simulated_data$BPMEDS ==
        1, simulated_data$SYSBP, 0)

    simulated_data$IS.framingham = 0

    test_set_temp <- test_set[, names(simulated_data)]

    simulated_data <- rbind(simulated_data, test_set_temp)

    return(simulated_data)
}
```

```
sim_men_1 <- sim_data(cor_matrix = cor_matrix_sim_1, std_vec = std_vec_men,
    mean_vec = mean_vec_men, p_vec = P_cate_men, nrow = nrow_men,
    test_set = test_set_men)
sim_men_2 <- sim_data(cor_matrix = cor_matrix_sim_2, std_vec = std_vec_men,
    mean_vec = mean_vec_men, p_vec = P_cate_men, nrow = nrow_men,
    test_set = test_set_men)
sim_men_3 <- sim_data(cor_matrix = cor_matrix_sim_3, std_vec = std_vec_men,
    mean_vec = mean_vec_men, p_vec = P_cate_men, nrow = nrow_men,
    test_set = test_set_men)
sim_men_4 <- sim_data(cor_matrix = cor_matrix_sim_4, std_vec = std_vec_men,
    mean_vec = mean_vec_men, p_vec = P_cate_men, nrow = nrow_men,
    test_set = test_set_men)

sim_women_1 <- sim_data(cor_matrix = cor_matrix_sim_1, std_vec = std_vec_women,
    mean_vec = mean_vec_women, p_vec = P_cate_women, nrow = nrow_women,
    test_set = test_set_women)
sim_women_2 <- sim_data(cor_matrix = cor_matrix_sim_2, std_vec = std_vec_women,
    mean_vec = mean_vec_women, p_vec = P_cate_women, nrow = nrow_women,
    test_set = test_set_women)
sim_women_3 <- sim_data(cor_matrix = cor_matrix_sim_3, std_vec = std_vec_women,
    mean_vec = mean_vec_women, p_vec = P_cate_women, nrow = nrow_women,
    test_set = test_set_women)
sim_women_4 <- sim_data(cor_matrix = cor_matrix_sim_4, std_vec = std_vec_women,
    mean_vec = mean_vec_women, p_vec = P_cate_women, nrow = nrow_women,
    test_set = test_set_women)

phi_beta_sim_men_1 <- Tran_Brier(sim_men_1, Y.fram.test.true.men,
    Y.fram.test.pred.men)
phi_beta_sim_men_2 <- Tran_Brier(sim_men_2, Y.fram.test.true.men,
    Y.fram.test.pred.men)
phi_beta_sim_men_3 <- Tran_Brier(sim_men_3, Y.fram.test.true.men,
    Y.fram.test.pred.men)
phi_beta_sim_men_4 <- Tran_Brier(sim_men_4, Y.fram.test.true.men,
    Y.fram.test.pred.men)

phi_beta_sim_women_1 <- Tran_Brier(sim_women_1, Y.fram.test.true.women,
    Y.fram.test.pred.women)
phi_beta_sim_women_2 <- Tran_Brier(sim_women_2, Y.fram.test.true.women,
    Y.fram.test.pred.women)
phi_beta_sim_women_3 <- Tran_Brier(sim_women_3, Y.fram.test.true.women,
    Y.fram.test.pred.women)
```

```r
phi_beta_sim_women_4 <- Tran_Brier(sim_women_4, Y.fram.test.true.women,
    Y.fram.test.pred.women)


# Create a data frame with the values
data <- data.frame(Men = c(phi_beta_sim_men_1, phi_beta_sim_men_2,
    phi_beta_sim_men_3, phi_beta_sim_men_4), Women = c(phi_beta_sim_women_1,
    phi_beta_sim_women_2, phi_beta_sim_women_3, phi_beta_sim_women_4))

rownames(data) <- paste0("Correlation setting: ", 1:4)
# Render the table using kable kable(data, format =
# 'latex', booktabs = TRUE, col.names = c('Male',
# 'Female')) %>% kable_styling(latex_options =
# c('hold_position')) %>% add_footnote(caption = 'Brier
# Scores of simulation data sets', footnote_as_chunk =
# TRUE, threeparttable = TRUE) Create a caption as a
# footnote
caption_text <- "Table 3: Brier Scores of simulation data sets"

# Render the table with a caption as a footnote using
# kableExtra
kable(data, format = "latex", booktabs = TRUE, col.names = c("Male",
    "Female")) %>%
    kable_styling(latex_options = c("hold_position")) %>%
    footnote(general = caption_text)
```