Evaluation of Cardiovascular Risk Prediction Models in NHANES Population

Liangkang Wang

2023-11-23

A brief summary (300 words or less) of the major results and conclusions aimed at a non-technical reader.

Introduction

- Overview of the project.
- Importance of evaluating prediction models in different populations.
- Brief description of the Framingham Heart Study and NHANES.

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.

```
                     Stratified by SEX
                      1              2              p      test
n                       4671           6190
CVD (mean (SD))         0.36 (0.48)    0.20 (0.40)  <0.001
SEX (mean (SD))         1.00 (0.00)    2.00 (0.00)  <0.001
TOTCHOL (mean (SD))   234.71 (42.51) 246.37 (46.48) <0.001
AGE (mean (SD))        54.24 (9.47)   54.85 (9.60)   0.001
SYSBP (mean (SD))     134.74 (20.01) 136.75 (24.16) <0.001
DIABP (mean (SD))      83.67 (11.14)  82.37 (11.78) <0.001
CURSMOKE (mean (SD))    0.51 (0.50)    0.37 (0.48)  <0.001
DIABETES (mean (SD))    0.05 (0.22)    0.04 (0.20)   0.007
BPMEDS (mean (SD))      0.06 (0.23)    0.11 (0.31)  <0.001
HDLC (mean (SD))       43.74 (13.30)  53.68 (15.88) <0.001
BMI (mean (SD))        26.26 (3.41)   25.61 (4.48)  <0.001
SYSBP_UT (mean (SD)) 125.46 (35.95) 119.47 (46.25) <0.001
SYSBP_T (mean (SD))     9.06 (37.16)  17.03 (50.13) <0.001
```

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.

```r
# blood pressure, demographic, bmi, smoking, and hypertension info
bpx_2017 <- nhanes("BPX_J") %>%
  select(SEQN, BPXSY1 ) %>%
  rename(SYSBP = BPXSY1)
demo_2017 <- nhanes("DEMO_J") %>%
  select(SEQN, RIAGENDR, RIDAGEYR) %>%
  rename(SEX = RIAGENDR, AGE = RIDAGEYR)
bmx_2017 <- nhanes("BMX_J") %>%
  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)) %>%
  select(SEQN, CURSMOKE)
bpq_2017 <- nhanes("BPQ_J") %>%
```

```r
  mutate(BPMEDS = ifelse(BPQ050A == 1, 1, 0)) %>%
  select(SEQN, BPMEDS)
tchol_2017 <- nhanes("TCHOL_J") %>%
  select(SEQN, LBXTC) %>%
  rename(TOTCHOL = LBXTC)
hdl_2017 <- nhanes("HDL_J") %>%
  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)) %>%
  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")

b <- CreateTableOne(data = df_2017, strata = c("SEX"))
table_matrix <- print(b, 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"))
```

Variable	Male	Female	p-value
n	4557	4697	
SEQN (mean (SD))	98363.83 (2677.38)	98296.19 (2665.73)	0.223
SYSBP (mean (SD))	122.49 (18.71)	120.20 (21.09)	<0.001
SEX (mean (SD))	1.00 (0.00)	2.00 (0.00)	<0.001
AGE (mean (SD))	34.12 (25.75)	34.55 (25.25)	0.419
BMI (mean (SD))	26.16 (7.63)	26.98 (8.80)	<0.001
HDLC (mean (SD))	49.57 (13.53)	57.01 (14.94)	<0.001
CURSMOKE (mean (SD))	0.21 (0.41)	0.14 (0.35)	<0.001
BPMEDS (mean (SD))	0.84 (0.37)	0.86 (0.35)	0.334
TOTCHOL (mean (SD))	176.68 (40.38)	182.94 (40.59)	<0.001
DIABETES (mean (SD))	0.11 (0.31)	0.09 (0.29)	0.001

```
# knitr::kable(table_matrix, format = "latex", booktabs = TRUE)
```

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:

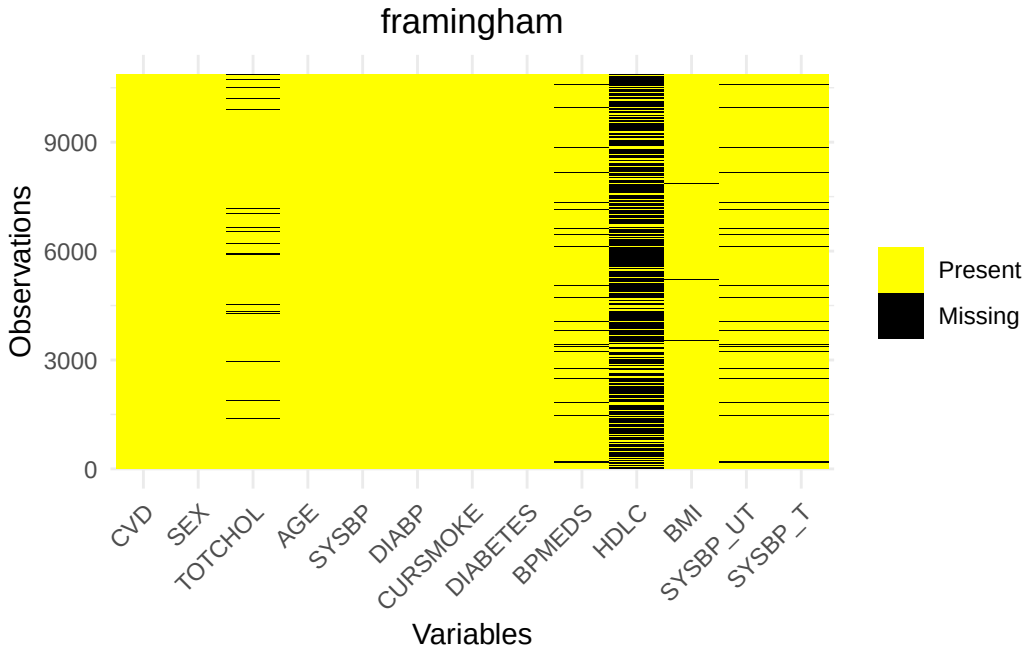
- `CVD`: Represents the occurrence of cardiovascular disease. It is a binary variable, where 1 indicates the presence of cardiovascular disease and 0 indicates its absence.
- `SEX`: Participant's sex, where 1 denotes male and 2 denotes female.
- `TOTCHOL`: Serum Total Cholesterol measured in mg/dL.
- `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).

Each of these variables plays a crucial role in our analysis, helping to elucidate the complex interplay of various risk factors in cardiovascular health.

Data Integration and Preprocessing

- Merging datasets.
- Handling missing data and outliers.



- Feature selection and engineering.

Methodology

Model Building

- Description of the predictive model.
- Model training and validation methods. ## Model Evaluation Metrics
- Explanation of evaluation metrics (e.g., Brier score, AUC). ## Transportability Analysis
- Techniques used to estimate model performance in the NHANES population.

Results

Model Performance in the Framingham Study

- Analysis of model accuracy and other metrics. ## Transported Model Performance in NHANES
- Comparison of model performance in the NHANES data. ## Simulation Study Results
- Findings from the simulated transportability analysis.

Discussion

- Interpretation of results.
- Comparison with existing literature.
- Implications for healthcare and predictive modeling.

Limitations

- Limitations of the methods used.
- Limitations of the data sources.

Conclusion

- Summary of key findings.
- Potential future research directions.

References

List of all references cited in the report.

Appendices

Additional Tables and Figures

- Supplementary material that supports the analysis. ## R Code
- Link to the GitHub repository containing the project code.