PHP2550 Project 3

Transportability of Prediction Models.

Background: Simulation studies are virtual experiments that involves generation of data through random sampling from known probability distributions (Morris, White, and Crowther 2019). These studies are key in modern research and analysis. Simulation studies provide a controlled and mimicked environment essential for investigating complex systems. They are widely employed in biostatistics to evaluate and compare different statistical methods under varied situations. These studies help analysts understand how statistical methods behave, as they involve knowing some truth about the parameters of interest during the data generation process (Morris, White, and Crowther 2019). Despite their significance,reviewers have previously shown that poorly designed and analyzed simulation studies often lead to the uncritical use and interpretation of results(Morris, White, and Crowther 2019). In that context, our study follows the ADEMP guidelines described by Morris (2019), aiming to address these challenges in this simulation study. Specifically, our simulation study aims to evaluate the performance of a predictive logistic risk model for cardiovascular heart disease, originally derived from the Framingham Heart Study data and subsequently transporting measures of model performance to estimate their performance within the population underlying the NHANES (National Health and Nutrition Examination Survey) survey data. By employing the ADEMP guidelines for the methods we achieve transparency in our simulation study and contribute to the understanding of the model's transportability and its potential application in diverse populations.More information about the sources of the data are publicly available at: https://www.cdc.gov/nchs/nhanes.htm and https://biolincc.nhlbi.nih.gov/studies/framcohort/.

Method: This study uses data from Framingham Heart Study to train predictive logistic risk models for predicting the risk of cardiovascular heart disease in men and women in the Nhanes population in the US. A weighted generalized logistic regression model was employed on the Framingham data to predict the risk of cardiovascular Heart disease with inverse estimated odds of source participation used as weights. The ADEMP principles was employed to conduct the simulation study with rigor and transparency, and to transport the prediction score from the Framingham study to a sample target population obtained from the Nhanes study.Data was generated based on the covariate information of the Nhanes sample based on the known probability and the specified population quantities. Nhanes data lacked long-term outcome information, therefore a comprehensive approach was utilized for integrating data from both NHANES and the Framingham Heart Study based on the variables that were common between these two populations. Eligibility criteria for the Framingham study to the Nhanes sample was also applied based on age and previous history of a stroke and heart attack as the inclusion criterion from the Framingham study. The integrated data was used to obtain inverse odds weights that were used in training the Framingham models and subsequent evaluations on target sample (Nhanes) as well as the simulated sample.The estimated weighted brier scores were presented for evaluating the models.

Results: Data with 2539 complete cases was used in training the Framingham predictive model and evaluated on the 2838 complete data of the Nhanes sample . On Nhanes data, the model achieved an weighted brier score of 0.19 and 0.11 on cardiovascular prediction on men and women respectively both on the simulated and non-simulated data. These scores were relatively low suggesting a moderate level of accuracy of predicting the probability of a cardiovascular heart disease and a good calibration. The model achieved an AUC of 0.7795 and 0.7248 for women and men respectively.

Conclusion: In this dynamic realm of study, simulations enhances our ability to make evidence-based decisions, particularly in scenarios where real world experiments may be impractical, unavailable or ethically challenging. The same weighted Brier scores of the prediction model in the target populations compared to

the source population reflects the major similarities in the covariate distribution between the two populations indicating that transportability analysis can be achieved given the scenarios where data is generated based on the available covariate information from the target population. The results shows that the model's performance is similar in Nhanes sample and simulated Nhanes population. The analysis provides insights into how well logistic regression models predict cardiovascular disease (CVD) in different populations.

Commonly used: National Health and Nutrition Examination Survey (NHANES)/target population and Framingham Heart Study (Framingham)/source population.

Introduction

Simulation studies mimics the real world and serve as an invaluable tool for understanding complex systems in a controlled and virtual environment. Application of simulation studies has risen in the recent past in various fields (Bienstock and Heuer 2022). Its application has significantly grown in research. Simulation studies enhances the understanding of the hypothetical scenarios and optimize strategies to be adopted in different fields such as predictive modelling especially in scenarios where obtaining real data seem challenging or where outcome data is missing. In this context the application of simulation studies provides a platform for risk assessment, decision making and further innovations in research. In the ever evolving world of technology and research, the application of simulation studies have increased greatly and have continued to be used to gain insights and making informed decisions across diverse domains(Boulesteix et al. 2020).

In recent years, several methods have been developed to evaluate the performance of prediction models in a target population and transporting measures of model performance from the source population to the target population(Steingrimsson et al. 2023).

Simulation studies have become integral to biostatistics. They serve as a tool for methodological evaluation, and the assessment of statistical models(Boulesteix, Strobl, and Augustin 2020). In this field, simulations are routinely employed to compare the performance of different statistical methods, prediction models and optimize clinical trial designs(Bienstock and Heuer 2022).

Users of prediction models are usually interested in using model-derived predictions in some target population (Steingrimsson et al. 2023). For instance, a healthcare system may employ predictive models for the likelihood of cardiovascular events in individuals receiving medical care, which may help help to identify individuals at a higher risk (Steingrimsson et al. 2023).In most cases, the source data (used in building the prediction model) are different from the target data as they drawn from different samples (Steingrimsson et al. 2023).

Our goal is therefore is to customize the Framingham model to the Nhanes population and to evaluate its performance in Nhanes population as well as the simulated population commonly referred to as transporting a prediction model.

This study is a collaboration with Dr. Jon Steingrimsson in the Biostatistics Department at Brown University. The study's overall gioal is to understand how well a prediction model perform in a target population that differs from the population where the model was developed and evaluated in. We evaluate the performance of cardiovascular risk prediction model in a target population underlying NHANES sample. We also conduct an evaluation where the target population is simulated based on the covariate information of the target population.

Data

The analysis used data from 2539 Framingham Heart (1094 men and 1445 women) and 2838 Nhanes samples who met Framingham eligibility criteria. The Framingham (source) data was used as the train set while the Nhanes (target) as the test set.

Missing Data

Table 1: Missing Data in Target Population

	Variable	Missing Count	% Proportion
SYSBP_UT	SYSBP_UT	505	17.79
SYSBP	SYSBP	448	15.79
HDLC	HDLC	306	10.78
TOTCHOL	TOTCHOL	306	10.78
SYSBP_T	SYSBP_T	276	9.73
BPMEDS	BPMEDS	184	6.48
BMI	BMI	163	5.74
DIABETES	DIABETES	1	0.04

As illustrated in Table 1, the observed missing data within the target population (NHANES) shows a non-random pattern, as indicated by varying proportions across different variables. Notably, the variable `SYSBP_UT` has the highest missing count at 17.79%, implying a potential systematic pattern in non response . Similarly, variables `SYSBP`, `HDLC`, and `TOTCHOL` also show relatively higher proportions of missing values, with `HDLC` and `TOTCHOL` missing together implying MNAR in the missing pattern.

In the analysis, we considered the potential impact of MNAR missingness on biasing estimates within the target population. To mitigate this, we employed the `mice` package in R for imputation. This method generated five imputed datasets, with complete estimates. The imputed datasets were then used in further analysis.

Methods

We assessed the sex specific calibration performance of a logistic regression model for prediction of cardiovascular disease (CVD) risk in Nhanes data. We used a composite data for the combined source and target populations based on common variables. We included an indicator variable `S` to distinguish between the source (Framingham data) and the NHANES datasets in the composite dataset. A logistic regression model was used to estimate the probability of membership in the source population. Using these model we derived the inverse-odds weights and applied these weights in training the predictive logistic regression model for CVD prediction using the Framingham sample. We then estimate the model performance in the NHANES target population. The model's performance was evaluated using weighted Brier scores and mean squared errors. The sex specific models showed slightly different score on predicting on Nhanes data. The model for men had a slightly higher Brier score (0.192) than the model for women (0.117).

Table 2: Measures of calibration on Nhanes Data

Metric	Model for Men	Model for Women
Brier Score	0.193	0.117
MSE Score	0.193	0.117

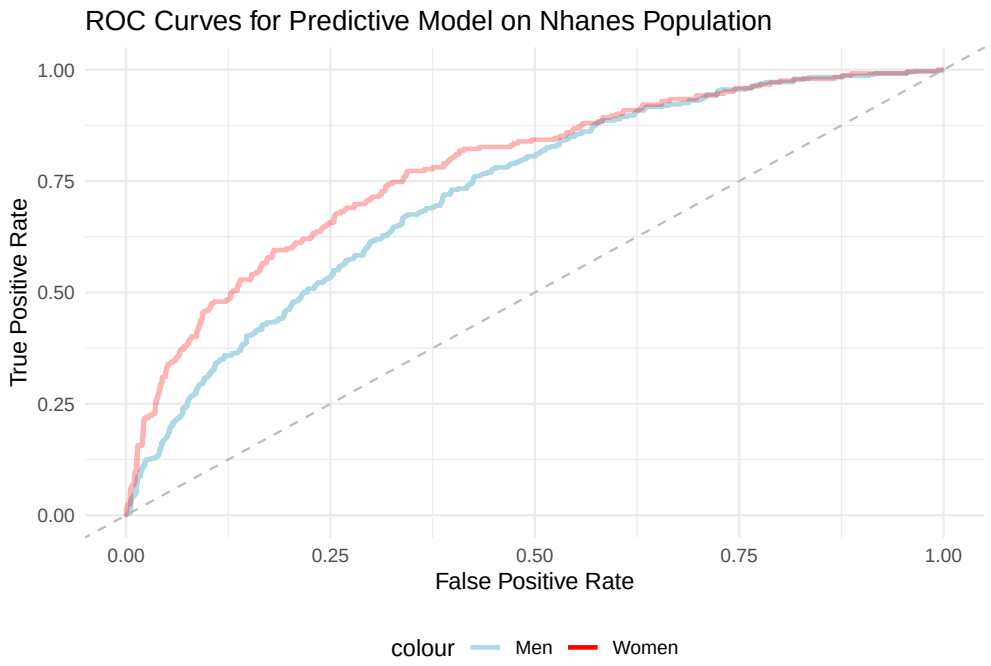

Figure 1: ROC Curve for the Model on Non-Simulated data

Prediction Logistic Models.

We adapted and assessed prediction models for Nhanes (target) population, that differed from Framingham (source) population where the sex-specific models were originally developed. We relied on using covariate information from the Nhanes population. Our approach for transporting prediction models assumed that the outcome and population were conditionally independent given the covariates

$$\frac{Pr[S=0|X, D_{\text{test}}=1]}{Pr[S=0|X, D_{\text{test}}=1]}$$

and the positivity of the likelihood of being in the Framingham population for each covariate pattern in the Nhanes population. We employed inverse-odds weights to apply the prediction model and evaluate its performance in the target population. This involved estimating the probability of belonging to the Framingham population based on covariates. We applied these methods to both simulated and real data for transporting a prediction model for Cardiovascular Heart disease from the Framingham Heart study to the US population of Nhanes eligible individuals.

Simulations Design

Individual level data was simulated for men and women separately based on the covariate information as seen in the summary statistics from Table 3 for each variable on the target population (Nhanes). The simulation size, was determined by the specified number of men $N = 1326$ and $N = 1512$ for women from the summary Table 3 so as to generate a pseudo-sample that resembles the Nhanes sample. We set two different random seed for each simulation to ensure reproducibility.The simulation included all the variables that were in the summary table. The continuous data (e.g., `SYSBP`, `AGE`, `BMI`, `HDLC`, `TOTCHOL`, `SYSBP_UT`, `SYSBP_T`) were simulated through a multivariate normal distribution with parameter values for mean and standard deviations set from the summary output in Table 3 and the binary data (e.g., `MCQ160E`, `MCQ160F`, `CURSMOKE`, `BPMEDS`, `DIABETES`) generated using a binomial distribution using the proportions from summary

table. These parameters were utilized to simulate data that closely resembled the characteristics of the target (Nhanes) population.

Simulation Results

As shown in Table 4 the simulated summary statistics closely match the summary statistics of the Nhanes dataset. Although some of the means and standard deviations of the simulated data are slightly different from the non-simulated sample which could be due to the variable transformation. The little overlap implies that the simulation's captured distribution assumptions and effectively reproduced the individual characteristics from the Nhanes dataset, indicating reliability of the simulated results in representing the underlying Nhanes data.

Table 3. Summary Statistics for Nhanes Data
Stratified by Sex

Characteristic	**1**, N = 1,326[1]	**2**, N = 1,512[1]
SYSBP	126 (16 72, 212)	122 (19 88, 224)
AGE	47 (10 30, 62)	46 (10 30, 62)
BMI	30 (7 16, 86)	31 (8 16, 75)
HDLC	48 (15 11, 166)	58 (16 18, 178)
CURSMOKE		
0	994 (75%)	1,265 (84%)
1	332 (25%)	247 (16%)
BPMEDS		
0	960 (78%)	1,119 (78%)
1	263 (22%)	312 (22%)
TOTCHOL	194 (40 84, 431)	195 (38 94, 352)
DIABETES		
0	1,177 (89%)	1,356 (90%)
1	149 (11%)	155 (10%)
MCQ160E		
2	1,326 (100%)	1,512 (100%)
MCQ160F		
2	1,326 (100%)	1,512 (100%)
SYSBP_UT	93 (54 0, 190)	88 (53 0, 192)
SYSBP_T	25 (52 0, 180)	25 (53 0, 224)

[1] Mean (SD Range); n (%)

Table 4. Summary Statistics Across Simulations
Stratified by Sex

Characteristic	**1**, N = 1,758,276[1]	**2**, N = 1,758,276[1]
SYSBP	126 (16 47, 202)	126 (16 47, 202)
AGE	47 (10 -1, 95)	47 (10 -1, 95)
BMI	30 (7 0, 64)	30 (7 0, 64)
HDLC	48 (15 -30, 118)	48 (15 -30, 118)
TOTCHOL	194 (40 -9, 388)	194 (40 -9, 388)
SYSBP_UT	93 (54 -180, 375)	93 (54 -180, 375)
SYSBP_T	25 (52 -247, 273)	25 (52 -247, 273)
MCQ160E		
1	1,758,276 (100%)	1,758,276 (100%)

MCQ160F		
1	1,758,276 (100%)	1,758,276 (100%)
CURSMOKE		
0	1,305,902 (74%)	1,305,902 (74%)
1	452,374 (26%)	452,374 (26%)
BPMEDS		
0	1,380,911 (79%)	1,380,911 (79%)
1	377,365 (21%)	377,365 (21%)
DIABETES		
0	1,561,588 (89%)	1,561,588 (89%)
1	196,688 (11%)	196,688 (11%)

[1] Mean (SD Range); n (%)

Table 5: Comparison for Measures of calibration

Metric	Model on Simulated	Model on Non Simulated
Brier Score	0.1926, 0.1166	0.1926479, 0.1165783

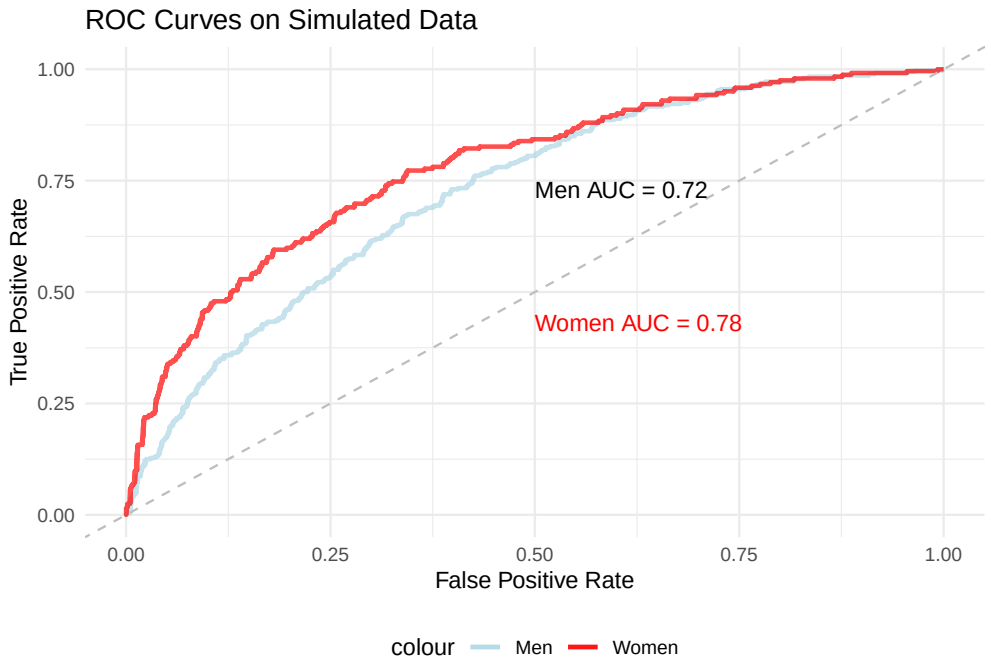

Figure 2: ROC Curve for Model on Simulated data

Transportability of Prediction Models.

In this study, we conducted a transportability analysis to assess the applicability of logistic regression models for predicting the likelihood of cardiovascular disease (CVD) in simulated Nhanes population. Using the Framingham Heart Study data as a source population, we simulated diverse datasets representing Nhanes covariate information. Logistic regression models were adapted for each simulation incorporating inverse-odds weights derived from the Framingham data. These weights, were computed based on predicted CVD

probabilities from the Framingham logistics model to adjust for the differences between the Framingham and Nhanes populations. The models were used to predict CVD probabilities in simulated data, and predictive accuracy was assessed using the Brier score.

Transportability Results

The Brier scores presented in Table 5 indicate comparable performance between the simulated and non-simulated data. Notably, the model predicting cardiovascular risk in men exhibited a slightly higher Brier score (0.1926) compared to the model for women (0.1166), suggesting a marginally lower predictive accuracy for cardiovascular risk in men as reflected by the higher Brier score.

Limitations

Our study had a number of limitations. First, the simulations were based on assumed distributions and parameters, which may have likely introduced uncertainty in the replication of real-world populations and does not consider outliers and extreme values. Secondly, the logistic regression models used in the study relied on a specific set of covariates, which could potentially be overlooking important variables not captured by the Framingham model.Additionally, the adaptation of inverse-odds weights assumed that the relationship between predictors and CVD remains consistent across populations, which may not hold in all instances.Furthermore, the application of the Framingham model to predict CVD probabilities in simulated population assumes that the underlying risk factors and relationships are comparable between genders.Despite these limitations, our analysis provides valuable insights into the transportability of the Framingham model and its performance, which underscores the complexities in applying predictive models across diverse populations especially where real data is lacking.

Conclusion

Our study found that the simulated and non-simulated data showed similar weighted Brier risk scores for predicting cardiovascular risk in Nhanes population.However, the model showed a slightly lower accuracy in predicting cardiovascular risk for men than for women. Overall, our predictive model works optimally when transported to different populations than the one it was built and evaluated in. This analysis contributes insights into the model's robustness and generalization when applied beyond its original study population. This work highlights the need for improvement of predictive models for cardiovascular outcomes.

References

Bienstock, Jared, and Albert Heuer. 2022. "A Review on the Evolution of Simulation-Based Training to Help Build a Safer Future." *Medicine* 101 (25): e29503. https://doi.org/10.1097/MD.0000000000029503.

Boulesteix, Anne-Laure, Rolf HH Groenwold, Michal Abrahamowicz, Harald Binder, Matthias Briel, Roman Hornung, Tim P Morris, Jörg Rahnenführer, and Willi Sauerbrei. 2020. "Introduction to Statistical Simulations in Health Research." *BMJ Open* 10. https://doi.org/10.1136/bmjopen-2020-039921.

Boulesteix, Anne-Laure, Carolin Strobl, and Thomas Augustin. 2020. "Introduction to Statistical Simulations in Health Research." *BMJ Open* 10 (12): e039921. https://doi.org/10.1136/bmjopen-2020-039921.

Morris, Tim P, Ian R White, and Michael J Crowther. 2019. "Using Simulation Studies to Evaluate Statistical Methods." *Statistics in Medicine* 38 (11): 2074–2102.

Steingrimsson, Jon A, Constantine Gatsonis, Bing Li, and Issa J Dahabreh. 2023. "Transporting a Prediction Model for Use in a New Target Population." *American Journal of Epidemiology* 192 (2): 296–304. https://doi.org/10.1093/aje/kwac128.

Code Appendix

```
# load required library
library(tidyverse)
library(kableExtra)
library(knitr)
library(pROC)
library(latex2exp)
library(ggplot2)
library(tidyr)
library(dplyr)
library(lattice)
library(riskCommunicator)
library(tableone)
library(nhanesA)
library(broom)
library(MASS)
library(gtsummary)
# load source data
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 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)

framingham_summary_stats <- 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)

# Filter to each sex
framingham_df_men <- framingham_df %>% filter(SEX == 1)
framingham_df_women <- framingham_df %>% filter(SEX == 2)

# Fit models with log transforms for all continuous variables
mod_men <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                log(SYSBP_T+1)+CURSMOKE+DIABETES,
```

```
      data= framingham_df_men, family= "binomial")


mod_women <- glm(CVD~log(HDLC)+log(TOTCHOL)+log(AGE)+log(SYSBP_UT+1)+
                   log(SYSBP_T+1)+CURSMOKE+DIABETES,
                 data= framingham_df_women, family= "binomial")


# 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)

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)

mcq_2017 <- nhanes("MCQ_J") %>%
  dplyr::select(SEQN, MCQ160E, MCQ160F)
```

```r
# Join data from different tables
nhanes_df <- 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") %>%
  full_join(mcq_2017, by = "SEQN") # added by keviner

# Eligibility criteria based on the framingham paper
nhanes_df <- nhanes_df %>%
  filter(AGE >= 30 & AGE <= 62) %>%
  filter(MCQ160E == 2 & MCQ160F == 2)

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

#nhanes_df$SEX <- as.factor(nhanes_df$SEX)
nhanes_df$CURSMOKE <- as.factor(nhanes_df$CURSMOKE)
nhanes_df$BPMEDS <- as.factor(nhanes_df$BPMEDS)

nhanes_df$MCQ160E <- as.factor(nhanes_df$MCQ160E)
nhanes_df$MCQ160F <- as.factor(nhanes_df$MCQ160F)
nhanes_df$DIABETES <- as.factor(nhanes_df$ DIABETES)

framingham_df$CURSMOKE <- as.factor(framingham_df$CURSMOKE)
framingham_df$BPMEDS <- as.factor(framingham_df$BPMEDS)
framingham_df$DIABETES <- as.factor(framingham_df$DIABETES)
#framingham_df$CVD <- as.factor(framingham_df$CVD)
#framingham_df$CVD <- factor(framingham_df$CVD, levels = c("0", "1"))

nhanes_summary_stats <- CreateTableOne(data = nhanes_df, strata = c("SEX"))


# Distribution of Missing Data for nhanes_df
missing_df <- data.frame(
  Variable = names(nhanes_df),
  missing_count = sapply(nhanes_df, function(x) sum(is.na(x)))
)

# Calculate percent missing
missing_df$percent_missing <- round(missing_df$missing_count / nrow(nhanes_df) * 100, 2)

# Arrange by percent missing in descending order
missing_df <- missing_df %>%
  arrange(desc(percent_missing))

# Select only those with missing records
```

```r
missing_df <- missing_df %>%
  filter(missing_count > 0)
missing_df$missing_count <- round(missing_df$missing_count, 2)
missing_df$percent_missing <- round(missing_df$percent_missing, 2)

missing_df %>%
kable(caption = "Missing Data in Target Population",
      col.names = c("Variable", "Missing Count", "% Proportion"),
        digits = 3,
        booktabs = TRUE) %>%
 kable_styling(latex_options = c("HOLD_position", "striped"),
               font_size=8)
```