STAC51 Final Project

Wenxin Gong, Kuei-Sheng Hou, Can Ming Jiang

2023-04-10

Risk Factors that Play Significant Roles in Osteoporotic Fracture
Group 7

Background and Significance, Discussion/Conclusion: Wenxin Gong:
Exploratory data analysis: Kuei-Sheng Hou:
Model: Can Ming Jiang

Word Count:

Library Used in Case Study:

library(ggplot2)
library("ggpubr")
library(dplyr)
library(corrplot)
library(ResourceSelection)
library(pROC)

Background and Significance

Abstact

Osteoporosis affects a significant number of people in the United States, with an estimated 54 million individuals affected. Fractures resulting from this condition are common, with up to one in two women and one in four men over 50 experiencing an osteoporosis-related fracture. This condition weakens bones and increases the risk of fractures in areas such as the spine, hip, wrist, and shoulder. The National Health and Nutrition Examination Survey (NHANES) collects data on risk factors such as age, gender, race, BMI, smoking, and alcohol use to identify underlying health conditions that contribute to osteoporotic fractures. This information is used to develop strategies for prevention and management. This case study uses data from the 2007-2008 NHANES survey, which included 10,149 individuals, to examine risk factors for osteoporotic fractures.

Significance of the Case Study

Despite being a national problem, osteoporosis often goes undetected and untreated. While NHANES has always provided many insights into the risk factors for osteoporosis, researchers have not exploited such information to the fullest. The current case study is significant because it ensures maximum data exploitation to identify the population's underlying risk factors for osteoporotic fracture. The case study is also significant because it looks at the predictors of osteoporotic fracture (i.e., fractures of the hip, wrist, and spine) in men and women and compares the results across different racial groups. This case study also brings to light the bone mineral density (BMD) measures that best predict osteoporotic fractures. The last significance of this case study is that it helps in developing effective strategies to prevent and manage osteoporosis. By examining the risk factors for fracture, including BMD measures after controlling for various non-BMD factors, this case study helps to develop the most effective strategies to prevent osteoporosis and reduce fracture risks. Therefore, by providing comprehensive information on fractures, this study will benefit policymakers, healthcare providers, and individuals in designing preventive strategies to reduce the burden of osteoporotic fractures.

Variable description

After cleaning the data, we have 17 variables to create a model

- ALQ130: average number of alcohol drinks/day in the last 12 months (numeric value with a maximum of 95; 777=refused; 999=don't know; .=missing)
- ALQ140Q: number of days having 5+ drinks in the last 12 months (numeric value with a maximum of 365; 777=refused; 999=don't know; .=missing)
- BMXBMI: body mass index (numeric value .=missing)
- DBQ229: regular milk use 5 times per week (1=a regular milk drinker for most or all of my life, including childhood; 2=never been a regular milk drinker; 3=milk drinking has varied over my life-sometimes I've been a regular milk drinker; 7=refused; 9=don't know; .=missing)
- DIQ010: doctor told you have diabetes (1=yes; 2=no; 3=borderline; 7=refused; 9=don't know; .=missing)
- MCQ190: type of arthritis (1=rheumatoid arthritis; 2=osteoarthritis; 3=other arthritis; 7=refused; 9=don't know; .=missing)
- DXXWDBMD: ward's triangle BMD (numeric value; .=missing)
- DXXL2BMD: L2 BMD (numeric value; .=missing)
- RIDAGEYR: age in years (numeric value with a maximum of 80; .=missing)
- WHD020: current self-report weight (pounds) (numeric value; 7777=refused; 9999=don't know; .=missing)

- MCQ160C: doctor ever said you had coronary heart disease (1=yes; 2=no; 7=refused; 9=don't know; .=missing)
- MCQ160L: ever told you had any live condition (1=yes; 2=no; 77777=refused; 99999=don't know; .=missing)
- OSQ130: ever taken prednisone or cortisone nearly every day for a month or longer (1=yes; 2=no; 7=refused; 9=don't know; .=missing)
- OSQ170: did mother ever fracture a hip (1=yes; 2=no; 7=refused; 9=don't know; .=missing)
- OSQ200: did father ever fracture a hip (1=yes; 2=no; 7=refused; 9=don't know; .=missing)
- RIAGENDR: gender (1=male; 2=female; .=missing)
- RIDETH1: race/ethnicity (1=Mexican American; 2=other Hispanic; 3=non-hispanic white; 4=non-hispanic black; 5=other race; .=missing)

Data Cleaning

Remove missing information and inaccurate variables.

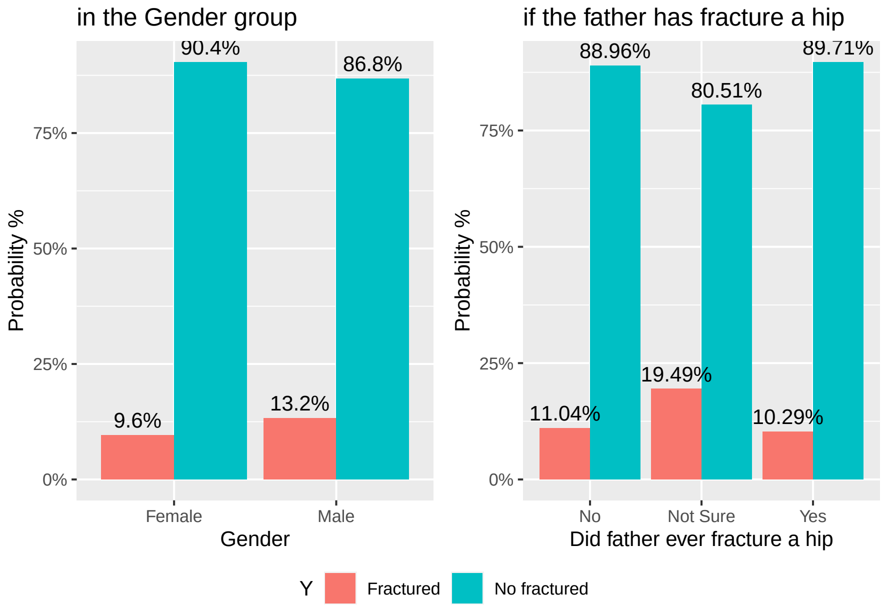
```
fracture.data = read.csv("Frax_Risk.csv")

D2 <- fracture.data[,!(names(fracture.data) %in%
                         c("SEQN", "DBQ197", "DIQ220", "MCQ160A", "MCQ180C",
                           "OSQ010A", "OSQ010B", "OSQ010C", "WHD110","MCQ180L",
                           "MCQ180A", "MCQ170L", "OSQ040AA", "OSQ040BA",
                           "OSQ040CA", "OSQ070"))]
D2 = D2[D2$DXXOFBMD > 0 & D2$DXXNKBMD > 0 & D2$DXXINBMD > 0 & D2$DXXOSBMD > 0 &
          D2$DXXTRBMD > 0 & D2$DXXWDBMD > 0 & D2$DXXL1BMD > 0 & D2$DXXL2BMD > 0 &
          D2$DXXL3BMD > 0 & D2$DXXL4BMD & D2$OSQ020B < 7777 & D2$ALQ130 < 777 &
          D2$WHD020 > 0 & D2$WHD020 < 7777,]
D2 = D2[D2$RIDAGEYR <= 80 & D2$RIDAGEYR > 0 & D2$BMXBMI > 0,]
D2[D2$MCQ190 == 0,]$MCQ190 = 9
D2[D2$OSQ170 == 7,]$OSQ170 = 9
D2[D2$OSQ200 == 7,]$OSQ200 = 9
```

Exploratory Data Analysis / Data Visualization

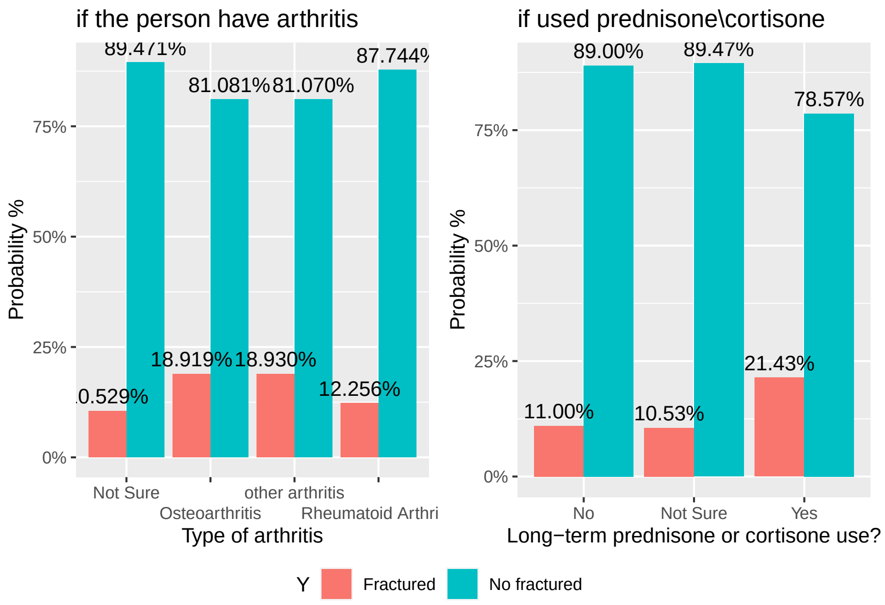
Categorical Variables that were used in the final model

What are the probability of getting fractured:



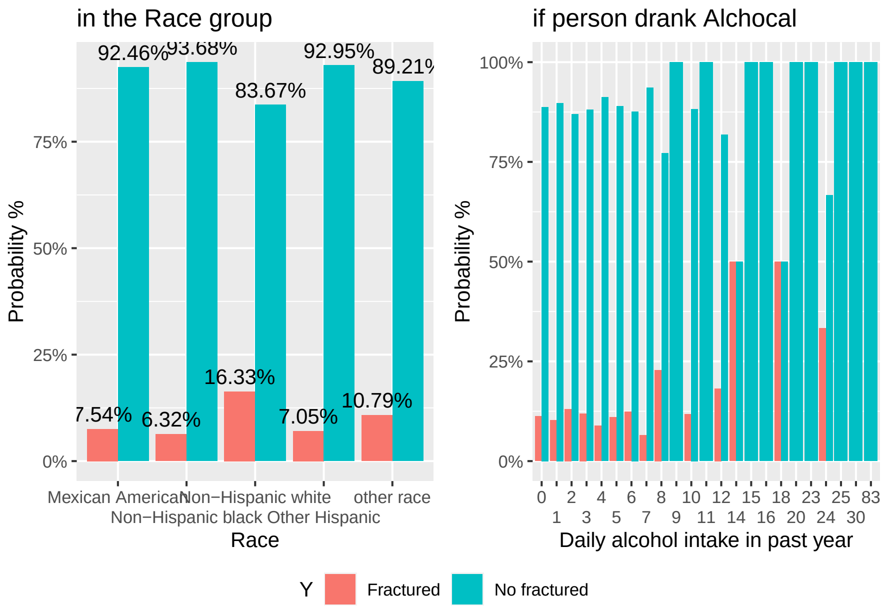
- In gender, male has a higher probability of getting fractured.
- From the graph, there is no clear evident that if the father has ever fractured a hip impacts the person, however, it it significant in the model. Interestingly, if the mother has ever fractured a hip is also a variable in the data set but it is not as significate compare to the father.

What are the probability of getting fractured:



- From the graph, Osteoarthritis and other type of arthritis has higher probability of getting fractured.
- Taking Prednisone or Cortisone long-term increase the probability of getting fractured.

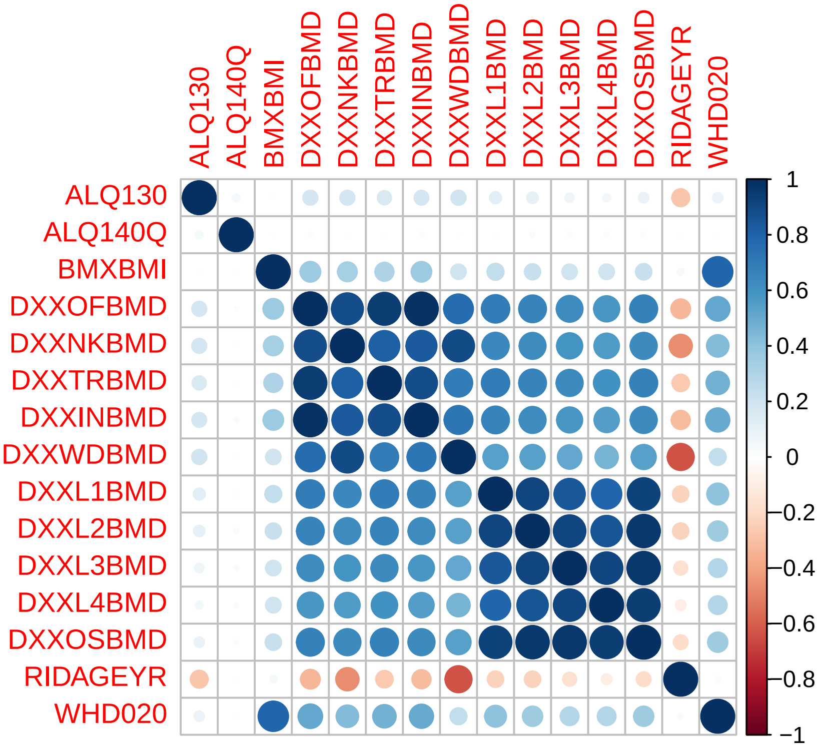
What are the probability of getting fractured:


in the Race group

if person drank Alchocal

- From the graph, non-Hispanic while has higher probability of getting fractured.
- Also, if the amount of daily alcohol drinks is high, it would also increase the probability of getting fractured

Check for MultiCollinearity



From the correlation plot, we can see that there are high correlation between DXXOFBMD, DXXNKBMD, DXXTRBMD, DXXINBMD and high correlation between DXXL1BMD, DXXL2BMD, DXXL3BMD, DXXL4BMD. We will select the variables in the Model Selection section.

Model Selection and Diagnosis

NOTE: model selection on report is done on testing data set, while results provided during presentation used full data set, So, some values may vary.

Cleaning The Data Set

After cleaning the data we have 17 variables to create a model. (ALQ130, ALQ140Q, BMXBMI, DXXWDBMD, DXXL2BMD, RIDAGEYR, WHD020, DBQ229, DIQ010, MCQ190, MCQ160C, MCQ160L, OSQ130, OSQ170, OSQ200, RIAGENDR, RIDETH1)

When fitting the model using the number of fractures(spine, hip, wrist), We initially used a poisson regression to model the data.However, we had issues creating a model that decreased the AIC. When examining the data one possible reason why the poisson regression did not work was that in the data set, there was no way to differentiate between 0 fractures or missing information. So, we switched over to using fracture or not as our response variable and applying a logistic regression. We also split the data set into two part one for training the model and one for testing the model. We decided on a 70 - 30 split between training and testing set.

```r
#Spliting data set into training(70% of data) and testing(30% of data)
set.seed(1234)
D2=cbind(fracture.y,fracture.n,D2)
sample=D2[sample(1:nrow(D2),nrow(D2)),]
training = sample[1:floor(0.7*nrow(D2)),]
testing = sample[(floor(0.7*nrow(D2))+1):nrow(D2),]
```

The variables DXXOFBMD, DXXNKBMD,DXXTRBMD,DXXINBMD,DXXWDBMD are highly correlated so it is sufficient to choose one when constructing the model, we will choose the one with the lowest AIC, in the single predictor models.

```r
fit.4 = glm(cbind(fracture.y,fracture.n)~DXXOFBMD,family=binomial,data=training)#AIC = 3722.2
fit.5 = glm(cbind(fracture.y,fracture.n)~DXXNKBMD,family=binomial,data=training)#AIC = 3709.3
fit.6 = glm(cbind(fracture.y,fracture.n)~DXXTRBMD,family=binomial,data=training)#AIC = 3721.5
fit.7 = glm(cbind(fracture.y,fracture.n)~DXXINBMD,family=binomial,data=training)#AIC = 3724.6
fit.8 = glm(cbind(fracture.y,fracture.n)~DXXWDBMD,family=binomial,data=training)#AIC = 3688.1
```

We choose to keep DXXWDBMD in the model since it has lowest AIC

Similarly we apply the same to the variables DXXL1BMD,DXXL2BMD,DXXL3BMD,DXXL4BMD,DXXOSBMD

```r
fit.9 = glm(cbind(fracture.y,fracture.n)~DXXL1BMD,family=binomial,data=training) #AIC = 3722.5
fit.10 = glm(cbind(fracture.y,fracture.n)~DXXL2BMD,family=binomial,data=training)#AIC = 3716.1
fit.11 = glm(cbind(fracture.y,fracture.n)~DXXL3BMD,family=binomial,data=training)#AIC = 3717.5
fit.12 = glm(cbind(fracture.y,fracture.n)~DXXL4BMD,family=binomial,data=training)#AIC = 3717.9
fit.13 = glm(cbind(fracture.y,fracture.n)~DXXOSBMD,family=binomial,data=training)#AIC = 3715.9
```

We choose to keep DXXOSBMD in the model since it has the lowest AIC

Due to the large number of variables we had to form the model, we were unable to form the saturated model and run it through the automated model selection processes. Therefore, we decided to first the model without interaction terms, then later add in interaction terms by hand, choosing the model which reduced AIC.

Now we will form the model without interaction terms

output for forward
Step: AIC=3566.46 cbind(fracture.y, fracture.n) ~ as.factor(RIDRETH1) + DXXWDBMD + as.factor(RIAGENDR) + as.factor(OSQ200) + as.factor(OSQ130) + DXXOSBMD + as.factor(MCQ190) + ALQ130 + WHD020

output for stepwise

Step: AIC=3566.46 cbind(fracture.y, fracture.n) ~ as.factor(RIDRETH1) + DXXWDBMD + as.factor(RIAGENDR) + as.factor(OSQ200) + as.factor(OSQ130) + DXXOSBMD + as.factor(MCQ190) + ALQ130 + WHD020

Output for backward
Step: AIC=3566.46 cbind(fracture.y, fracture.n) ~ ALQ130 + DXXWDBMD + DXXOSBMD + WHD020 + as.factor(MCQ190) + as.factor(OSQ130) + as.factor(OSQ200) + as.factor(RIAGENDR) + as.factor(RIDRETH1)

The methods forward, backward, stepwise of automated model selection all gave the same results.

Note, that since we did not want our model to be flooded with too many interaction terms, we made sure to limit the number of interaction terms during manual model selection.

After adding interaction terms by hand, AIC of our training model = 3559.3 and our model is cbind(fracture.y,fracture.n)~as.factor(RIDRETH1) + DXXWDBMD + as.factor(OSQ130) + as.factor(RIAGENDR) + DXXOSBMD+ RIDAGEYR + as.factor(RIAGENDR):as.factor(RIDRETH1) +DXXWDBMD:DXXOSBMD +ALQ130 +ALQ130:DXXOSBMD +RIDAGEYR:as.factor(RIAGENDR)

comparing training model with model without interaction terms

```
anova(fit.forward,train,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(fracture.y, fracture.n) ~ as.factor(RIDRETH1) + DXXWDBMD +
##     as.factor(RIAGENDR) + as.factor(OSQ200) + as.factor(OSQ130) +
##     DXXOSBMD + as.factor(MCQ190) + ALQ130 + WHD020
## Model 2: cbind(fracture.y, fracture.n) ~ as.factor(RIDRETH1) + DXXWDBMD +
##     as.factor(OSQ130) + as.factor(RIAGENDR) + DXXOSBMD + RIDAGEYR +
##     as.factor(RIAGENDR):as.factor(RIDRETH1) + DXXWDBMD:DXXOSBMD +
##     ALQ130 + ALQ130:DXXOSBMD + RIDAGEYR:as.factor(RIAGENDR)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      5216     3532.5
## 2      5214     3521.3  2   11.183  0.00373 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05, so the simplier model does not fit the data as well as the more complex model, So, our training model is superior to the on without interaction terms.

```
drop1(train, test="Chisq")
```

```
## Single term deletions
##
## Model:
## cbind(fracture.y, fracture.n) ~ as.factor(RIDRETH1) + DXXWDBMD +
##     as.factor(OSQ130) + as.factor(RIAGENDR) + DXXOSBMD + RIDAGEYR +
##     as.factor(RIAGENDR):as.factor(RIDRETH1) + DXXWDBMD:DXXOSBMD +
##     ALQ130 + ALQ130:DXXOSBMD + RIDAGEYR:as.factor(RIAGENDR)
##                                         Df Deviance    AIC     LRT  Pr(>Chi)
## <none>                                      3521.3 3559.3
## as.factor(OSQ130)                        2   3534.9 3568.9 13.6588 0.0010815
## as.factor(RIDRETH1):as.factor(RIAGENDR)  4   3538.4 3568.4 17.0891 0.0018574
## DXXWDBMD:DXXOSBMD                        1   3528.9 3564.9  7.5737 0.0059225
## DXXOSBMD:ALQ130                          1   3525.9 3561.9  4.5730 0.0324806
## as.factor(RIAGENDR):RIDAGEYR             1   3534.5 3570.5 13.1750 0.0002837
##
```

```
## <none>
## as.factor(OSQ130)                         **
## as.factor(RIDRETH1):as.factor(RIAGENDR) **
## DXXWDBMD:DXXOSBMD                         **
## DXXOSBMD:ALQ130                           *
## as.factor(RIAGENDR):RIDAGEYR              ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that all of the terms in our training model are significant. Also, by applying the Wald test from the summary() output, we also see that we are unable to drop any terms from our model.

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(train$y, fitted(train),g=21)#for training data
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  train$y, fitted(train)
## X-squared = 23.529, df = 19, p-value = 0.2148
```

```
hoslem.test(test$y, fitted(test),g=21)#for testing data
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  test$y, fitted(test)
## X-squared = 13.327, df = 19, p-value = 0.8214
```

When applying the Hosmer and Lemeshow goodness of fit test(because we have ungrouped data), we get that our model fits the data well since the p-value > 0.05.

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
par(mfrow=c(2,2))
test.roc = roc(train$y~fitted(train),plot=TRUE,print.auc=TRUE)#for training data
```

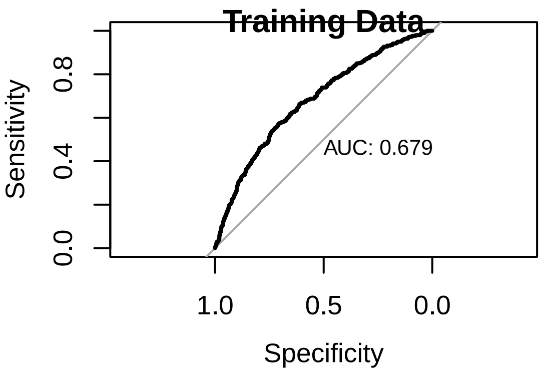
```
## Setting levels: control = 0, case = 1
```

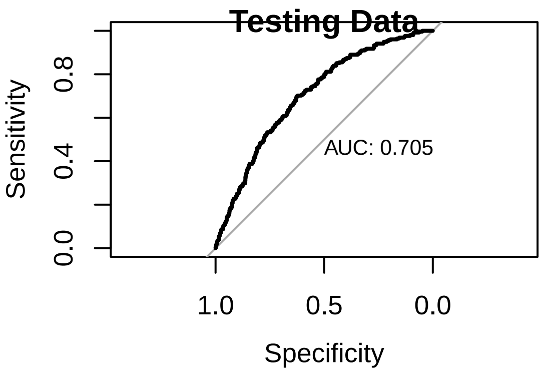
```
## Setting direction: controls < cases
```

```
title("Training Data")
test.roc = roc(test$y~fitted(test),plot=TRUE,print.auc=TRUE)#for testing data
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
title("Testing Data")
```





Classification Table

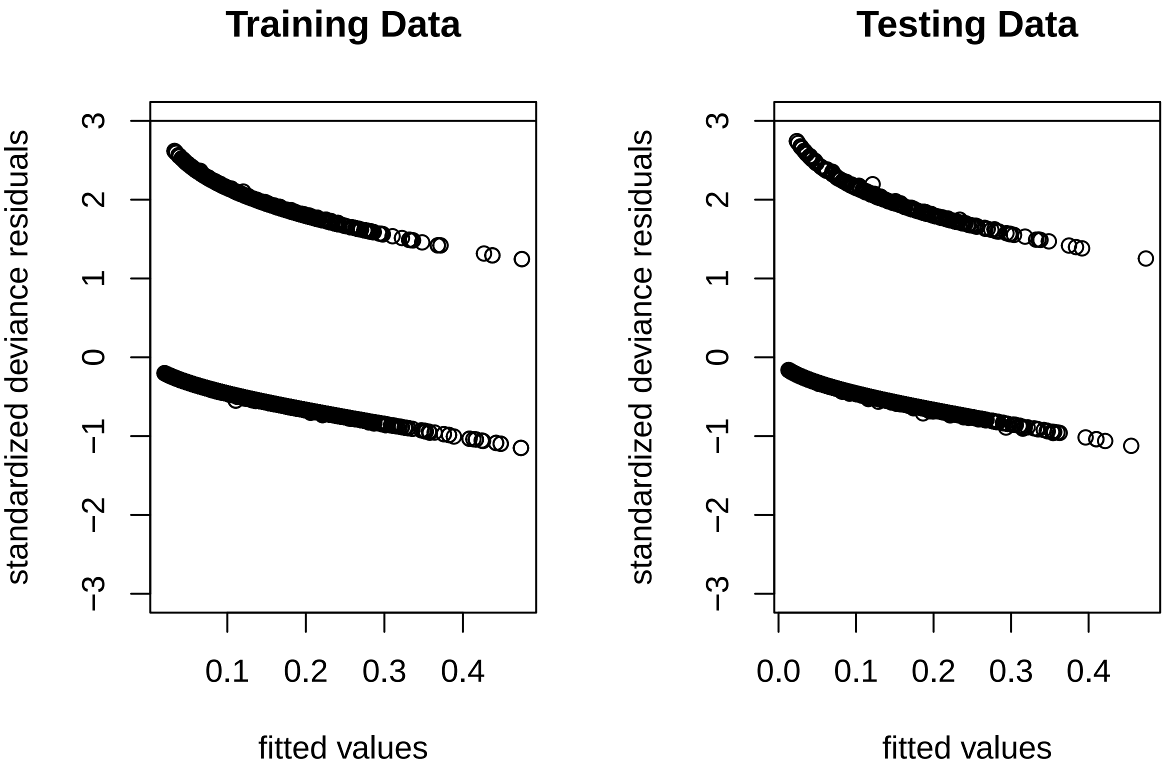
```
##        predicted.train
## y.train    0    1
##       0 2896 1735
##       1  221  381
```

```
##       predicted.test
## y.test    0    1
##      0 1203  786
##      1   76  179
```

```
##              data
## result         training   testing
##   sensitivity 0.6253509 0.6048265
##   specificity 0.6328904 0.7019608
```

We note that ideally we would like the area under our ROC curve as well as sensitivity and specificity to be higher however, rather than fitting a model with many terms, we decided to balance fitting a model that fits the data well and is useful as well as trying to minimize the number of parameters.

Perform Residual Analysis to check for outliers



Notice, that we have fitted values with residuals that have absolute values of less than 3 so, there is no evidence of outliers for both training and testing data sets.

Now for the other objectives, we will simply fit the model without interaction terms to get a general idea of which predictors are significant.

No interaction model for male data

It gives the following output Step: AIC=2793.37 fracture.m ~ as.factor(RIDRETH1) + DXXOSBMD + as.factor(OSQ200) + WHD020 + as.factor(OSQ130) + as.factor(DIQ010) + as.factor(MCQ160C) + ALQ130 + as.factor(MCQ190)

No interaction model for female data

It gives the following output Step: AIC=2204.09 fracture.f ~ DXXWDBMD + as.factor(RIDRETH1) + as.factor(MCQ190) + as.factor(OSQ130) + ALQ140Q + as.factor(DBQ229) + DXXOSBMD + as.factor(MCQ160C) + as.factor(OSQ200)

No interaction model for hip fracture or not

It gives the following output Step: AIC=5044.71 fracture.hip ~ as.factor(RIDRETH1) + DXXWDBMD + as.factor(RIAGENDR) + as.factor(OSQ200) + as.factor(MCQ190) + DXXOSBMD + as.factor(OSQ130) + ALQ130 + WHD020 + RIDAGEYR

Discussion/Conclusion

Our research question is to determine what are the risk factors for osteoporotic fracture. The primary objective of this case study was to identify predictors of osteoporotic fracture in men and women, focusing on fractures of the hip, wrist, and spine. The secondary objectives aimed at determining the best bone mineral density (BMD) measure(s) as a predictor of osteoporotic and hip fractures after controlling for various non-BMD risk factors. Our analysis revealed several significant predictors of osteoporotic fracture: ethnicity, spine BMD, ward's triangle BMD, gender, age and alcohol consumption. The interaction terms in the model, such as as.factor(RIAGENDR):as.factor(RIDRETH1), DXXWDBMD:DXXOSBMD, ALQ130:DXXOSBMD, and RIDAGEYR:as.factor(RIAGENDR), also indicate that the relationships between these factors and the risk of osteoporotic fracture are more complex and may depend on combinations of these factors. To be more specific, the results of the models without interaction terms showed that the significant predictors for men included ethnicity, spine BMD, family history of hip fracture, weight, prednisone or cortisone taken, history of diabetes, history of coronary heart disease, alcohol consumption and type of arthritis. For women, the significant predictors were ward's triangle BMD, ethnicity, type of arthritis, prednisone or cortisone taken, alcohol consumption, milk consumption, spine BMD, history of coronary heart disease and family history of hip fracture. For the secondary objectives, our analysis found that both ward's triangle BMD and spine BMD are important predictors of hip fracture risk when controlling for non-BMD factors. However, BMD measurements for predicting osteoporotic fracture need to be discussed in terms of gender differences. For men, spine BMD appears to be the best predictor, while for women, both ward's triangle BMD and spine BMD are important predictors. Overall, our findings suggest that a combination of demographic, lifestyle, and BMD-related factors can be used to predict the risk of osteoporotic fracture in men and women. These results can be used to identify individuals at high risk for osteoporotic fracture and to develop targeted prevention strategies. However, it is important to note that this study has several limitations. First, the data set used in this study was limited to a specific population and may not be generalizable to other populations. Second, the adaptation study relied on self-reported data, which may be subject to recall bias. Third, there may be other factors for inclusion in this analysis. Future research could build on this study by examining other risk factors for osteoporotic fracture and further exploring the relationship between BMD measurements and fracture. In addition, effective interventions for people at risk for osteoporotic fractures could be further investigated.

Reference

National Osteoporosis Foundation. (n.d.). Osteoporosis fast facts. Retrieved March 29, 2023, from http://www.bonehealthandosteoporosis.org/wp-content/uploads/2015/12/Osteoporosis-Fast-Facts.pdf

Centers for Disease Control and Prevention. (n.d.). Nhanes 2007-2008: Osteoporosis Data Documentation, codebook, and frequencies. Centers for Disease Control and Prevention. Retrieved April 9, 2023, from https://wwwn.cdc.gov/Nchs/Nhanes/2007-2008/OSQ_E.htm